AD 738299

COMPARATIVE EVALUATION OF THE RETRIEVAL EFFECTIVENESS OF DESCRIPTOR AND FREE-TEXT SEARCH SYSTEMS USING CIRCOL (CENTRAL INFORMATION REFERENCE AND CONTROL-ON-LINE)

Donald W. King
Peggy W. Neel
Barbara L. Wood

Westat Research, Inc.

Approved for public release;
distribution unlimited.

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)

Westat Research, Inc.
11600 Nebel St.
Rockville, MD 20852

2a. REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b. GROUP

3 REPORT TITLE

COMPARATIVE EVALUATION OF THE RETRIEVAL EFFECTIVENESS OF DESCRIPTOR AND FREE-TEXT SEARCH SYSTEMS USING CIRCOL (CENTRAL INFORMATION REFERENCE AND CONTROL ON-LINE).

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Final Report 1 April 1970 - 1 July 1971

5 AUTHOR(S) (First name, middle initial, last name)

Donald W. King
Peggy W. Neel
Barbara L. Wood

6. REPORT DATE

January 1972

7a. TOTAL NO. OF PAGES

126

7b. NO. OF REFS

5

8a. CONTRACT OR GRANT NO.

F30602-70-C-0205

b. Job Order Number: 45940000

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

0199

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

RADC-TR-71-311

10 DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11 SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Rome Air Development Center (IRDT)
Griffiss Air Force Base, New York 13440

13. ABSTRACT

·This study compares the retrieval effectiveness of two alternative input and search systems in terms of such measures as recall, fallout, precision, and total retrieval. One system operates using manually indexed document files searched by controlled vocabulary while the other employs full-text input using natural language searching. Both systems are applied to a common data base and hardware. Operational information needs were used in the form of request statements from actual users. From these statements of need, search queries were formulated for both systems and recall estimates calculated using a recall base that was pre-specified by the request originator. The queries were processed and total retrieval, fallout and precision ratios were calculated for both systems. The results indicate that the two systems perform at approximately the same level of effectiveness, although estimated average total retrieval was found to be slightly greater for free-text searching than for descriptor searching at all levels of recall. The primary conclusion from this study is that descriptor searching and free-text searching, as applied to CIRCOL, are sufficiently similar in terms of effectiveness as to necessitate some other basis for decision-making concerning the two systems. In addition to comparison of search systems, further evaluation concerned effects to be expected from various file changes, the relative performance of experienced and inexperienced users, analysis of recall failures under both systems, and cost effectiveness considerations using a system simulation model.

UNCLASSIFIED
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
DOCUMENTATION						
INFORMATION RETRIEVAL						
EVALUATION						
INFORMATION RETRIEVAL EFFECTIVENESS						

[illegible]--Griffiss AFB NY

UNCLASSIFIED
Security Classification

FOREWORD

This report was prepared by Westat Research, Inc., 11600 Nebel Street, Rockville, Maryland 20852, under Contract F30602-70-C-0205, Job Order No. 45940000. Research data were collected between August 1970 and June 1971.

The kind assistance of F. W. Lancaster, who acted as consultant for this project, was greatly appreciated. The authors also wish to thank all of the staff members of Foreign Technology Division, Wright-Patterson Air Force Base for their excellent cooperation in conducting this study. In particular, we wish to thank Mr. Charles Mangio, Dr. James Canfield and Mr. James Shawley for their leadership and cooperation. Mr. Nicholas M. Di Fondi (IRDT) was the RADC Project Engineer.

The study concerns an evaluation of the effectiveness of Central Information Reference and Control On-Line (CIRCOL). Also contained herein is a generalized cost/effectiveness model for more universal evaluation of document storage and retrieval systems. The basic data for model effectiveness parameters were found under this contract for Descriptor and Free-Text Search Systems using CIRCOL. These data were incorporated into the model along with effectiveness data results observed in other studies. Cost information was gathered from a number of sources under another contract and do not reflect costs of CIRCOL.

This technical report has been reviewed by the Office of Information (OI) and is releasable to the National Techncal Information Service (NTIS).

This technical report has been reviewed and is approved.

Approved: Nicholas M. Difondi
NICHOLAS M. DIFONDI
Technical Evaluator

Approved:
FRANZ H. DETTMER
Colonel, USAF
Chief, Intel & Recon Division

FOR THE COMMANDER: Fred I Diamond
FRED I. DIAMOND
Actg. Chief
Plans Office

ABSTRACT

This study compares the retrieval effectiveness of two alternative input and search systems in terms of such measures as recall, fallout, precision, and total retrieval. One system operates using manually indexed document files searched by controlled vocabulary while the other employs full-text input using natural language searching. Both systems are applied to a common database and hardware. Operational information needs were used in the form of request statements from actual users. From these statements of need, search queries were formulated for both systems and recall estimates calculated using a recall base that was prespecified by the request originator. The queries were processed and total retrieval, fall-out and precision ratios were calculated for both systems. The'results indicate that the two systems perform at approximately the same level of effectiveness, although estimated average total was found to be slightly greater for free-text searching than for descriptor searching at all levels of recall. The primary conclusion from this study is that descriptor searching and free-text searching, as applied to CIRCOL, are sufficiently similar in terms of effectiveness as to necessitate some other basis for decision-making concerning the two systems. In addition to comparison of search systems, further evaluation concerned effects to be expected from various file changes, the relative performance of experienced and inexperienced users, analysis of recall failures under both systems, and cost/effectiveness considerations using a system simulation model.

EVALUATION

The objective of this study was to evaluate the retrieval effectiveness of the Central Information Reference and Control On-Line (CIRCOL) document reference retrieval system when free-text generated index files are used as document representations as opposed to manual index files. Retrieval effectiveness is reported as the averages of Recall (the proportion of the total number of relevant documents that are retrieved), Precision (the proportion of the retrieved documents that are relevant), Fallout (the proportion of the total number of non-relevant documents that are retrieved), and total number of documents retrieved. This report includes: an initial analysis which was used to establish the experimental design for the main evaluation and to formulate hypotheses for a failure analysis; the results of the main evaluation; the results of a failure analysis of the components of retrieval; and a cost effectiveness model that can be applied to any retrieval system to determine the cost incurred per relevant document retrieval.

A significant conclusion derived from the initial and main evaluations is that neither type of indexing is clearly the best in terms of retrieval effectiveness. Any decision to continue CIRCOL with either type of indexing must be based on system goals. For example, free-text indexing provides the advantage of greater flexibility in the retrieval process because it does not restrict query formulation to a thesaurus or to specific word forms and it does allow the use of synonyms. This advantage should increase the Recall level of operation but at the expense of larger total retrievals. Manual indexing is not flexible and is completely dependent on the quality of the indexing which results in lower Recall levels of operation with lower total retrievals. Advantages and disadvantages such as these must be weighed against system goals in order to determine which is the appropriate indexing scheme for CIRCOL.

A significant conclusion derived from the failure analysis is that only in a very small number of cases are the indexing schemes at fault for the loss of relevant documents in retrieval. A great proportion of the failures were due to the user who requests documentation through a statement of need but fails to accurately describe his interest, or the system analyst who formulates a query from such statements of need but does not use an appropriate search strategy to effect accurate retrieval.

As a result of this study, future research can be directed toward investigation into file structures, user aids, interactive on-line capability, and other techniques to improve the CIRCOL operation.

Nicholas M Di Fondi

NICHOLAS M. DIFONDI

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

Appendix Page

LIST OF ILLUSTRATIONS

Figure

TABLE OF CONTENTS (continued)

LIST OF TABLES

TABLE OF CONTENTS (continued)

SECTION I

INTRODUCTION

The purpose of this study is to compare retrieval effectiveness of two input and search systems. These are:

- A manual indexing system based on humanly assigned index terms selected from a thesaurus. The search file thus created can be searched on-line or off-line using Boolean-type search equations.
- A free-text searching system using the natural language of the complete input document or of a portion of it, e.g., an abstract.

With searching software Boolean logic can call out word combinations in document texts in on-line or off-line modes. Word roots (stem-suffix cutoff) can also be searched, as can consecutive strings of characters whereever they appear embedded in the text. Word proximity (co-occurrence in a paragraph or sentence, for example) can be used as a syntactic substitute to improve precision in searching.

An additional purpose of the study is to provide a general framework for the evaluation of these types of retrieval systems.

For several reasons, the operational system of the Foreign Technology Division (FTD) of Wright-Patterson Air Force Base was chosen for comparison of the two retrieval systems. The FTD system contains a large file of references and a small part of this database is already common to both systems under consideration. FTD management had expressed an interest in a comparative analysis of the two systems.

The great attraction of free-text searching is that it reduces the cost of indexing. Manual indexing using a controlled vocabulary is an expensive operation creating its own problems including the maintenance of consistency and dependence upon skilled and trained personnel who are difficult to recruit and retain. The maintenance of a dynamic, frequently updated controlled vocabulary for such an indexing operation is also an expensive proposition. Free-language searching may offer certain economies in input operations. This is particularly true for documents that can be obtained in machine-readable form as a byproduct of another operation.

While comparison of retrieval systems may be carried out in several ways, two basic measures are common. These are the recall ratio or proportion of relevant documents retrieved and the corresponding number of documents retrieved. Other measures may be combined with these in order to conduct more meaningful analysis.

The Central Information Reference and Control (CIRC) at the FTD of Wright-Patterson Air Force Base, Ohio, formed the basis for the first CIRCOL (Central Information Reference and Control On-Line) database. CIRC was designed to provide intelligence analysts of the scientific and technical community with bibliographic information. At one time processing of requests was carried out using an IBM 7094 in a batch mode. Later CIRCOL processing was conducted using an IBM 360/65 computer with manually indexed files and IBM Document Processing System (DPS) software. At this time only FTD analysts had access to the system. More recently, in addition to the manual indexing of all incoming documents, a large portion of these have been processed to allow free-text on-line searching of abstracts along with the assigned topic tags or descriptors. Thus, there is an increasing body of documents that is searched by means of abstracts. CIRCOL users now include intelligence analysts from a variety of government agencies. Both the number of analysts and the number of searches conducted vary considerably between agencies and over periods of time.

At the outset of the study the CIRCOL database contained approximately 365,000 documents. When data collection for this study was completed the CIRCOL database had grown to approximately 533,000 documents. The initial 365,000 documents of the file were processed by manual indexing, and that portion of the file thus includes only assigned topic tags and titles for each document. The remaining 168,000 document references composing the CIRCOL database are a mixture of titles, topic tags and abstracts. As incoming documents were processed beyond the first 365,000 an increasing proportion of the total file included searchable abstracts as well as topic tags and titles.

The subject matter of CIRCOL, as described in FTD's Users' Guide, includes seven types of document information:

- STEP (Scientific and Technical Exploitation Program) Information Subsystem (SIS) inputs, which provide information from Communist country sources concerning research and development in the aerospace sciences and technologies. STEP is a function of the Aerospace Technology Division, Library of Congress.

- MIS (Miscellaneous Inputs Information Subsystem), which processes documents, articles, and abstracts from various open-source literature and some classified sources. Major sources include Chemical Abstracts, JPRS abstracts, FBIS summaries and FTD summaries. This program is a function of Project Have Stork.

- IRIS (Intelligence Reports Information Subsystem), which is designed to input selected raw intelligence reports received by FTD from Army, Air Force, Navy, DIA and other sources. Intelligence reports are screened for FTD interest and those selected are processed into CIRCOL.
- ITIS (Internal Translation Information Subsystem), which is FTD human and machine translations.
- IFRIS (Intelligence Finished Reports Information Subsystem), which includes studies, reports, etc. resulting from a project or task funded by FTD. This is composed of foreign technology scientific and technical documents produced by the Foreign Technology Division, the Deputies for Foreign Technology of the AFSC Divisions and Centers, and their contractors.
- EFRIS (External Finished Reports Information Subsystem), designed to process reports which fall in FTD's area of interest and are produced by elements external to AFSC foreign technology organizations and their contractors.
- BAIS (Bulletin Articles Information Subsystem), containing technical briefs cn several scientific and technical disciplines which are published periodically in the FTD Bulletin.

Rapid growth of the CIRC database necessitated standardization of the terminology. CIRC has seen a transition from essentially uncontrolled indexing to partial control to the rigid control of a thesaurus.

All incoming documents are assigned topic tags from the CIRC Thesaurus, which is composed of Official Terms, Synonyms, and Official Nomenclature Terms. It is made up of three volumes, which are:

- Subject-structural vocabulary
- Permuted vocabulary
- Alphabetized vocabulary

Also included as part of the Thesaurus are references to Scope Notes, Broader Terms, Narrower Terms, Synonyms and See Also terms.

CIRCOL searches are run on two types of terminals (IBM 2471's and AT&T or WW models 33 and 35). Initial searching options available to users are controlled vocabulary terms, authors, country codes, and free-text searching of available abstracts. A user may choose to limit the number of references from his search by using one or more of the following qualifying elements: date, country of information, type of information, subject area, classification, publishing country, update information and accession number.

Output format may be selected by the user from a list of 13 options. Output of the first 365,000 documents of the database will include only bibliographic information, topic tags and title. When requested off-line, however, the remainder of the file will also include those abstracts that are available in the system. The CIRCOL communication network is entirely unclassified and references to classified documents are made through a bibliographic entry. Microfilm files are available at each terminal for those documents without abstracts available through the system.

In this report a "search" is defined as a particular information need and a "query" or "search query" as the terms and the associated logic developed to fulfill an information need or search. Searching on CIRCOL may be done either by the technical analyst or by a system monitor or search analyst. For purposes of this study written statements of information requirements based on actual need were solicited by the CIRCOL search analyst from FTD's technical analysts. Initial search queries were developed by the search analyst and the output judged for relevance to the previous written statement of need by both the search analyst and the technical analyst separately, using abstracts obtained from the microfilm file. Subsequently, new queries for these statements were developed and run by Westat personnel. The technical analysts also were asked to list as many documents as possible prior to conducting the search to be used as a sample recall base for later estimations. This information formed the basis for most of the study. Subsets of information collected are described within the appropriate parts of this report.

One of the measures used was an estimate of recall ratio, which, as explained above, is the proportion of relevant documents retrieved to relevant documents in the file. Another measure was the estimated fallout ratio, which is the proportion of nonrelevant documents retrieved. A third measure concerned the precision ratio, or the proportion of documents retrieved that were relevant. The fourth measure was the retrieval size, or total retrieval (number of documents retrieved by a search query).

In order to accomplish the stated objectives of the evaluation, experimentation was divided into two phases. This was necessary partly because of the extremely small size of the body of the database containing both descriptors and abstracts. However, a two-phase approach appeared to be advisable in any circumstances. Phase 1 was designed to be primarily a diagnostic study and pretest using the basic measures of recall, fallout, precision, and total retrieval in order to gain a clearer understanding of the systems in question and to identify hypotheses that could be tested during Phase 2. In addition to the further testing of hypotheses developed during Phase 1, Phase 2 consisted of additional in-depth evaluation of the parameters and environments surrounding the two major systems in question. Included were such factors as the effect to be expected from the use of relevant documents during query formulation, the effect that system experience has upon search results, and an analysis of various proposed file changes. As an additional feature a cost/effectiveness model was provided for use by FTD in future decision-making.

SECTION II

IMPLICATIONS OF THE STUDY

Some of the more important implications from the study include:

- Higher recall levels appear to be achieved under free-text searching than under descriptor searching at an increase in number of documents retrieved for operational searches.
- When search sequences are performed over specified levels of recall the descriptor searches appear to yield slightly fewer number of documents retrieved at all levels of recall.
- A controlled vocabulary may be useful but is not needed for full-text searching.
- Recall failures stem more from search procedures and user/system interface problems than from input sources under both free-text and descriptor searching.
- Using known relevant documents during query formulation appears to improve search effectiveness.
- Improved search effectiveness can be expected from users with greater system experience.
- Various word-file changes yield increased retrieval size without corresponding improvement in recall level.

In measuring and comparing the retrieval effectiveness provided by manual indexing and free-text processing, parameters such as recall, retrieval size,precision, and fallout, are helpful in evaluation of similarities and differences between the two systems. Other factors, however, must also be considered prior to any ultimate decisions or formulation of final conclusions concerning the two systems.

Phase 1 of this study attempted to compare the effectiveness of the following alternative systems in terms of the four measures listed above:

(1) Actual Descriptor System - Controlled vocabulary queries were formulated from the analyst's statement of need and searched on topic tags.

(2) Ideal Descriptor System - Controlled vocabulary queries were formulated using prespecified relevant documents and searched on topic tags.

(3) Full-Text Controlled Vocabulary System - Controlled vocabulary queries were formulated using the analyst's statement of need and searched on full-text.

(4) Full-Text Natural Language System based on Need-Natural language queries were formulated using the analyst's statement of need and searched on full-text.

(5) Full-Text Natural Language System based on Query-Natural language queries were formulated using the Actual Descriptor System queries and searched on full-text.

(6) Full-Text Natural Language System based on Ideal-Natural language queries were formulated using the recall base of prespecified relevant documents and searched on full-text.

A summary of the results is shown below for thirty searches.

Effectiveness categories	Descriptor		Full-Text Controlled Vocabulary	Full-Text search based on		
	Actual 1	Ideal 2	3	Need 4	Query 5	Ideal 6
Recall (%)	43	82	40	66	65	84
Retrieval	46	156	30	118	86	86
Precision (%)	41	23	60	25	34	43
Fallout (%)	.0074	.0329	.0033	.0244	.0156	.0134

Both of the Ideal Systems (2, 6) were included to demonstrate the margin for potential improvement of the particular system and were therefore excluded from further evaluation, although the Full-Text System appears to be best under ideal circumstances. Since recall is about the same but total retrieval is less for full-text searching, the Full-Text Controlled Vocabulary System (3) was included in the test to determine if a controlled vocabulary might be needed even for full-text searching. Results indicate that controlled vocabulary searches on full-text (3) may be slightly better than those on topic tags (1). However, natural language searches (4) from the same statements of need yield improved recall with increased total retrieval. Inclusion of the Query System (5) was an attempt to isolate the effect of the controlled vocabulary on system performance, whether operating under a descriptor or full-text system.

Of the six systems, the two that were of most concern were the Descriptor System (i.e., manually indexed files) and the Full-Text Natural Language System (i.e., free-text processing). The results of Phase 1 analysis indicated that for the same set of requests the Descriptor System obtained an average recall of 43 percent, with precision of 41 percent and retrieval size of 46 documents, while the Full-Text Natural Language System (free-text processing) achieved an average recall of 66 percent, with precision of 25 percent and retrieval size of 118. On this basis, the decision as to which system provided the best performance or effectiveness depended on a choice among recall, total retrieval or precision. The Descriptor System provided higher precision and smaller retrieval size, both of which are desirable. However, the recall was lower. On the other hand, although recall was higher for the Full-Text System, the precision was lower and the retrieval size larger, both of which are undesirable. Thus, when considering recall in isolation, the Full-Text System looks best, although with regard to precision and retrieval size the Descriptor System appears to be superior. The final conclusion based on these three parameters would, of course, depend upon the objectives of system users and consequently the weight that is given to each factor.

It was this uncertainty in the comparison and evaluation that played a major role in the decision to design Phase 2 as a two-system comparison over four predefined recall levels (25 percent, 50 percent, 75 percent, 100 percent). This was accomplished through broadening a sequence of search queries. The results of this four-level comparison are shown below.

Retrieval size

Recall level	25 percent	50 percent	75 percent	100 percent
Descriptor System	32	72	140	252
Full-Text System	38	86	157	296

Precision (%)

Recall level	25 percent	50 percent	75 percent	100 percent
Descriptor System	41	36	28	21
Full-Text System	34	30	25	18

Fallout (%)

Recall level	25 percent	50 percent	75 percent	100 percent
Descriptor System	.0052	.0126	.0028	.0548
Full-Text System	.0068	.0164	.0032	.0669

From these results, it appears that the Descriptor System outperforms the Full-Text System at all levels of recall although by a relatively small magnitude. The sample size of the test was not sufficiently large to distinguish between the two results at reasonable levels of statistical significance.

One point that is not revealed in such a comparison, however, involves the typical level of operation of both systems in terms of recall, fallout, precision, and retrieval size. Since the same requests were involved in the formulation of queries for both Phase 1 and Phase 2 systems the average performance of each system may be viewed as a typical level of operation. This would indicate the possibility of achieving, in normal operation, higher levels of recall in the Full-Text System than in the Descriptor System. If the Full-Text System consistently provides higher levels of recall the benefits may, in terms of lower precision, larger fallout and larger retrieval size, outweigh the greater cost. Compensation may also accrue from the additional time made available to the analyst because of the increased ease and speed of query formulation without use of a thesaurus. In short, the number of possible ways of retrieving a particular document is much greater with a free-text system than with a descriptor system. Under a descriptor system if a document is improperly indexed the error is difficult to correct. The burden is placed on the indexer under an index system and on the searcher under free-text systems. A descriptor system also allows the possibility of losing a document for all practical purposes through faulty indexing. A document may be indexed in a certain manner due to the current emphasis on or importance of the subject matter or area of discussion or because of limitations imposed by the level of exhaustivity required by the indexing policy. At a later date if the importance of the document shifts to a previously obscure portion, the indexing will not allow retrieval while a similar case under the Full-Text System would still be retrievable by searching additional terms in the text. This process is facilitated in an on-line system.

Recall failures were investigated for both the Descriptor System and the Full-Text System. In the Descriptor System 27.5 percent of the recall failures were attributable to index language and indexing process while 49 percent of the failures were due to searching process and 23.5 percent due to the user/system interaction. In the Full-Text System 12 percent of the recall failures were due to synonym problems while 38 percent were caused by searching process and 50 percent attributable to user/system interface.

One other possible system design considered briefly in this study was a combination of searchable descriptors and abstracts such as is currently being used for incoming CIROL documents. An analysis of a small sample of (20 searches) revealed that with the combination design no changes occurred in recall, but the average retrieval size increased from 17 documents in a regular full-text system searching only titles and abstracts to 22 documents for a system searching titles, abstracts, and topic tags.

SECTION III

RESULTS OF THE STUDY[1]

This study was divided into two phases. The first was primarily a diagnostic study to develop hypotheses to be tested in Phase 2. The second phase included more operational experimentation and compared two of the systems from Phase 1 analysis in more detail along with various facto : s concerning the system environments.

1. Phase 1

a. Introduction

Three systems were chosen for initial evaluation.

- Indexed input with controlled vocabulary descriptor retrieval
- Full-Text (e.g., abstracts, extracts, etc.) input with controlled vocabulary retrieval
- Full-Text input with natural language retrieval

The first system was approximately the same as that employed by CIRCOL. The second system utilized the same controlled vocabulary search strategy as that of System 1, with matching (and associated retrieval), however, on the basis of the abstract (including title if applicable) instead of descriptors. This last system was evaluated by using a three-part breakdown on the basis of search strategy formulation.

b. Phase 1 Results

Within the three major systems six subsystems were evaluated in Phase 1 in terms of two primary measures which were recall or proportion of relevant documents retrieved and number of documents retrieved.

Briefly, the search queries used under System 1 were those formulated by the CIRCOL search analyst according to regular procedures, using a controlled vocabulary and descriptor searching. Those used in System 2 were based on descriptors chosen through examination of known relevant documents. This system thus represented an ideal-type System 1. System 3 queries were exactly the same as those used under System 1 (controlled vocabulary); however, retrieval was determined by free-text

[1] Details of calculations, methods used, and examples may be found in Appendices I and II.

searching of abstracts instead of descriptor searching as in Systems 1 and 2. above. Queries for Systems 4, 5, and 6 were all formulated using natural language terms and retrieval was determined under all three systems by free-text searching of abstracts rather than descriptor searching. The differences among the last three systems (4, 5, and 6) rested on the basis used for query formulation. System 4 (Need) queries were formulated from the intelligence analyst's statement of need (Form A) and therefore represented the free-text searching equivalent of System 1 (Descriptor). System 5 (Query) queries were formulated from the System 1 queries in order to isolate and include the effect of the controlled vocabulary itself. System 6 (Ideal) queries were developed as System 2 queries were (e.g., by examination of the known relevant documents from Form B) and therefore represented an ideal-type System 4. More detailed discussion of the design and methodology is included in Appendix I.

The results for these six systems are shown in the following two-by-two tables.

Actual Descriptor System

	Relevant	Not relevant	
Retrieved	19	27	46
Not retrieved	25	364,929	364.954
	44	364,956	365,000

Ideal Descriptor System

	Relevant	Not relevant	
Retrieved	36	120	156
Not retrieved	8	364,836	364,844
	44	364,956	365,000

Full-Text Controlled Vocabulary System

	Relevant	Not relevant	
Retrieved	18	12	30
Not retrieved	26	364,944	364,970
	44	364,956	365,000

Full-Text Natural Language Systems

		Relevant	Not relevant	
Need	Retrieved	29	89	118
	Not retrieved	15	364,867	364,882
		44	364,956	365,000

		Relevant	Not relevant	
Query	Retrieved	29	57	86
	Not retrieved	15	364,899	364,914
		44	364,956	365,000

		Relevant	Not relevant	
Ideal	Retrieved	37	49	86
	Not retrieved	7	364,907	364,914
		44	364,956	365,000

Table I shows a summary of these results.

Table I Summary of effectiveness figures for the six Phase 1 systems

Effectiveness categories	Descriptor		Full-Text Controlled Vocabulary	Full-Text search based on		
	Actual 1	Ideal 2	3	Need 4	Query 5	Ideal 6
Average recall (%)	43	82	40	66	65	84
Average number documents retrieved	46	156	30	118	86	86
Precision (%)	41	23	60	25	34	43
Fallout (%)	.0074	.0329	.0033	.0244	.0156	.0134

Tables II and III show details for the 30 searches.

Table II Recall ratios for Phase 1 searches by system

Search	Descriptor		Full-Text Controlled	Full-Text search based on		
	Actual	Ideal	Vocabulary	Need	Query	Ideal
	1	2	3	4	5	6
1	1/5	3/5	1/5	0/5	2/5	5/5
2	1/6	6/6	2/6	2/6	4/6	6/6
3	1/4	2/4	1/4	4/4	3/4	4/4
4	5/5	5/5	5/5	3/5	5/5	5/5
5	0/6	3/6	1/6	4/6	4/6	5/6
6	0/4	3/4	0/4	4/4	4/4	4/4
7	0/5	4/5	0/4	3/5	1/4	3/5
8	2/5	5/5	1/5	5/5	4/5	5/5
9	3/10	8/10	2/10	6/10	7/10	7/10
10	0/2	2/2	0/2	0/2	0/2	2/2
11	0/2	2/2	0/2	0/2	0/2	2/2
12	11/11	11/11	11/11	11/11	11/11	11/11
13	5/6	6/6	4/6	6/6	6/6	6/6
14	0/5	2/5	0/5	2/5	2/5	4/5
15	9/9	9/9	9/9	9/9	9/9	9/9
16	0/3	2/3	1/3	2/3	1/3	2/3
17	1/4	3/4	1/4	2/4	2/4	3/4
18	0/10	5/10	0/10	3/10	3/10	4/10
19	3/5	3/5	3/5	2/5	3/5	3/5
20	0/9	9/9	0/9	7/9	2/9	7/9
21	9/10	9/10	10/10	10/10	10/10	10/10
22	1/7	7/7	0/7	0/7	3/7	6/7
23	2/3	3/3	1/3	1/3	1/3	2/3
24	5/7	6/7	5/7	5/7	5/7	5/7
25	0/5	5/5	2/5	2/5	2/5	4/5
26	0/4	3/4	0/4	2/4	2/4	3/4
27	3/5	4/5	0/7	7/7	7/7	7/7
28	4/5	4/5	3/5	4/5	4/5	3/5
29	7/8	7/8	7/8	8/8	7/8	8/8
30	3/7	5/7	2/7	4/7	2/7	5/7
TOTAL	76/177	146/177	72/178	118/179	116/178	150/179
Recall average (%)	43	82	40	66	65	84
Standard Error (%)	8.1	4.5	8.3	6.3	5.9	4.4

Table III Total adjusted retrieval for Phase 1 searches by system based on a 365,000-document database

Search	Descriptor		Full-Text Controlled Vocabulary	Full-Text search based on		
	Actual	Ideal		Need	Query	Ideal
	1	2	3	4	5	6
1	5	126	3	119	50	36
2	32	168	3	11	4	250
3	6	55	17	205	16	199
4	15	14	3	1	69	0
5	27	220	8	1	13	14
6	2	382	14	264	191	210
7	6	468	1	18	4	77
8	52	52	27	2	325	2
9	20	82	10	76	128	145
10	18	35	9	114	16	42
11	4	1	2	24	30	0
12	20	9	10	228	17	19
13	28	55	43	68	50	55
14	69	13	36	16	11	12
15	34	34	18	2	30	3
16	135	112	102	493	171	252
17	8	8	4	7	12	5
18	19	59	10	32	23	41
19	34	8	18	12	30	20
20	44	312	23	113	103	90
21	44	31	37	77	38	26
22	59	442	31	71	157	104
23	182	7	95	14	162	7
24	38	472	20	48	47	58
25	5	279	3	221	7	178
26	13	209	22	406	294	109
27	46	59	24	180	59	49
28	170	300	88	250	181	101
29	163	335	152	479	254	420
30	92	341	48	24	80	62
Total retrieval	1,390	4,688	881	3,576	2,572	2,586
Average retr.	46	156	30	118	86	86
Standard error	9.4	-	-	21.6	-	-

Certain assumptions were made in analyzing the six systems in terms of recall and total retrieval. It was assumed that the two major objectives to be considered were the maximization of recall and the minimization of total retrieval (number of documents retrieved per search). Naturally, in combining the two objectives it is necessary to make various tradeoff decisions in terms of the basic parameters and goals of CIRCOL personnel, both at the management and analyst levels.

Table I shows that CIRCOL (Actual Descriptor System)operated at a level of 43 percent recall (on the average) with approximately 46 documents retrieved per search. The Ideal Descriptor System (2) indicated that there was the possibility of improving recall under a descriptor system to a level of approximately 82 percent, with an associated increase in the number of documents retrieved to 156 per search. Thus, improvement in recall might be brought about by having several known relevant documents made available to the search analyst either prior to or during the search process. For an evaluation of this procedure refer to Section III 2.d. This is indicated since System 2 queries were formulated using relevant documents as both a guide or pointer to additional terms which should be included, and as an indication of how well the proposed query would perform in terms of expected recall level. Another possible advantage of using relevant documents prior to the search may lie in the further ability of the analyst (when an intermediary is being used) to interpret the information need at hand. In other words, a written request may fail to express the true need of one person to another. It should be noted that the recall improvement from 43 percent to 82 percent indicated the potential for improvement of the present Descriptor System. The use of relevant documents in query formulation was tested in Phase 2 of the study. In System 3, it appears that using the same procedure for query formulation as that employed by CIRCOL (controlled vocabulary) and performing this search by free-text searching of abstracts (instead of descriptor searching) resulted in the achievement of approximately the same level of recall as that formerly attained (40 percent as compared to 43 percent) with an associated drop in the average number of documents retrieved per search (30 as opposed to 46 documents). Thus it appeared that merely changing the search procedure from indexed descriptor to full-text free-text searching, while holding the controlled vocabulary constant resulted in an improvement in size of retrieval with a very slight, if any, degradation in recall level. This was thought to have been the case since many of the terms occurring in the controlled vocabulary did not occur in exactly the same form or sequence in the full-text document. It would appear that System 3 was more highly discriminating (i.e., higher precision) than System 1 in that the recall levels were approximately equivalent while the number of documents retrieved by System 3 was considerably less than System 1. This means that most of those documents that were not retrieved by System 3 and were by System 1 were not relevant. Another possibility that appeared was that System 3 was retrieving different relevant documents from those retrieved by System 1. Upon examination of the experimental data, however, it appeared that the retrieved relevant documents were largely the same ones.

When considering Systems 4, 5, and 6, it should first be noted that all three involved full-text input as opposed to descriptor indexing and the use of natural language as opposed to the previous controlled vocabulary. As far as practicality is concerned, System 4 would be the most likely operational system of these three. It corresponded to the CIRCOL method of search formulation (based on a written request) except using natural language instead of a controlled vocabulary and free-text searching in place of descriptor searching. Under this System (4) the recall level was increased to 66 percent with a corresponding increase in the total retrieval size per search to 118 documents. System 5 differed from System 4 in that the query was formulated on the basis of the System 1 descriptor search developed by the CIRCOL search analyst. System 5 showed a slight decrease in recall and a substantital decrease in average total retrieval. This could have been due to not fully understanding the process that takes place during the search analyst's (intermediary's) translation of a written request into a controlled vocabulary query. The choice of terms or perhaps the nature of terms available in the vocabulary from which they were chosen seemed to cause a decrease in total retrieval even when these terms were translated into natural language. System 5 was included merely for comparative purposes and appeared not to be the most practical method of system operation. However, it should not be completely disregarded in light of the indicated level of recall and total retrieval. System 6 and its associated recall and total retrieval levels of 84 percent, 86 documents respectively represented the potential improvement possible for System 4 (written request, natural language query, and free-text searching of abstracts). As with System 1 and System 2, one possible method of system improvement may be the use of known relevant documents in search query formulation for the same reasons discussed previously. System 6 results also indicated that the possibility exists not only for improving (increasing) recall under System 4 but for decreasing the number of documents retrieved per search as well. Various changes in search strategy may be developed to aid in this improvement. The use of relevant documents is such a method.

Concentrating attention on the three basic systems of 1, 3 and 4, which one was considered to be better than the other two depended upon the parameters and goals of CIRCOL. For instance, assuming that great emphasis is placed on obtaining the highest possible proportion of relevant documents, then System 4 would be chosen without regard to the relatively high number of documents retrieved. On the other hand if only a moderate proportion of relevant documents is necessary, System 3 might be chosen because of its low level of total retrieval. If both are given moderate weights, then

System 1 might well prove to be desirable since both recall and total retrieval levels are relatively moderate values.

Considering the practice of using known relevant documents in order to improve search capability when an intermediary is used, it would appear to be advantageous from a cost/effectiveness standpoint to utilize this procedure in certain cases while not in others. This assumption is made without access to CIRCOL cost information; however, the cost/effectiveness model in Section IV demonstrates the procedure for making such a determination. A relatively easy method of determining the need for this procedure is to have the intelligence analyst specify the level of recall his particular search need requires along with his request. Normal search methods would then be employed in all cases except those requiring high recall levels. In the latter case the intelligence analyst would be asked to identify several relevant documents to be used by the intermediary in conducting the search. If the intelligence analyst is unable to identify any relevant documents the search analyst could then make a preliminary search and forward a sample of the output to the intelligence analyst in order to aid in the identification of relevant documents. This process should yield the desired results. This procedure is designed for use in cases when an intermediary is employed to conduct the search. It may also prove to be beneficial, however, to use basically the same concept to improve recall when the intelligence analyst himself is conducting the search. By examining several known relevant documents it is often possible to identify additional terms which had previously been overlooked and also to focus in on certain concepts which were chosen during the indexing process. Frequently, the same document may be viewed differently by indexers and searchers. Examining several known relevant documents can, if nothing more, serve to assure the analyst that his search design is progressing as planned and is following the correct path.

At the conclusion of Phase 1 certain hypotheses were formed concerning the failure analysis portion of the study. System failures resulting in the nonretrieval of relevant documents may be categorized into three groups:

- Indexing failures
- Failures of the search process
- Failures occurring in user/system interaction

Typical errors were, of course, different for the Descriptor System and the Full-Text Natural Language System.

Several areas of possible failure were hypothesized concerning the Descriptor System input. One of the largest sources of failure appeared to lie in the choice of index terms. Other terms were chosen which were mentioned directly in the document when concepts contained therein were overlooked. Another area of possible failure was created by the constraints imposed by the vocabulary ınvolved and the level of exhaustivity practical in indexing. An additional source of failure was brought about by the particular hierarchical nature of the vocabulary. Since narrower terms are not necessarily contained in broader terms, the burden to choose accurately the level of indexing is difficult for the searcher. Often it was merely guesswork as to what level of specificity would yield desired references. No pattern could be found for the choice of a general rather than a specific term or vice versa. Also the indexing policy followed (noninclusion of narrower terms in broader terms) produced cases in which the broader term carries fewer postings than the narrower term.

When considering the Full-Text Natural Language System it was obvious that failures would not include group one (indexing) sources of error. Most failures resulted from either the search process or user/system interaction. It was felt at this point that most problems did not concern the latter choice. The largest source of full-text failure generally mentioned in the literature involves synonym problems although there may have been some difficulty caused in this system as well by the hierarchical nature of language. There are many methods of overcoming a synonym problem, among them being use of synonym or related term list generation (either manual or automatic) and associative retrieval. An automatic cross-referral list or capability would serve the same purpose, as would a thesaurus. The burden of synonym choice in a full-text natural language system usually rests with the searcher. The choice of leaving this burden to the searcher may prove to be the best in terms of cost/effectiveness decision-making

One suggestion for improving the effectiveness and ease of use of the Full-Text System concerns the output display. Fiequently, a particular group of abstracts on a printout is extremely long and it proves to be a tedious task to identify portions of interest. A method of overcoming this is to highlight search terms by using all capitals when these terms appear in each abstract or title. These terms may be further highlighted by being printed in the margin beside each line of type which contains the term.

Two capabilities currently not in use which may prove to be extremely helpful in raising the efficiency of the Full-Text System are positional modifiers. The capability of searching for a combination of terms within the same sentence or paragraph instead of merely in a +1 or +2 relationship may prove vital primarily for improving precision. Another factor which may be important is the degree of accessibility of the full text for screening purposes, etc.

2. Phase 2

Phase 2 of the experimental analysis concerned the evaluation of the two following systems:

- Descriptor or Index System
- Full-Text Natural Language System

Each part of the Phase 2 Design was included in order to analyze some specific aspect of these systems.

Part A allowed the performance (in terms of recall, total retrieval, fallout, and precision) of each system to be compared with that of the other system. This comparison was displayed graphically.

Part B concerned an effectiveness model that enables each system to be viewed in terms of several alternatives such as employing or not employing an intermediary (search analyst). A cost model was provided that will allow FTD to compare the cost/effectiveness of the two systems under examination.

Part C involved the area of recall failure and categorized these failures into four groups. This provided a basis for further recommendations for system improvement.

Part D provided the opportunity to analyze the effects of both the extent of technical analyst experience with the system and the use of known relevant documents on system performance during the search process.

Part E investigated possible effects of adding such files as Word Form Conversion and Synonym Equivalent on system performance.

For Standard Error estimates of these parts refer to Appendix I.

a. Part A - System Comparison Over Four Levels of Recall

In order to compare the Descriptor System and the Full-Text Natural Language System, queries were developed which would retrieve the four desired levels of recall (25 percent, 50 percent, 75 percent, 100 percent) for each of the two systems. This was done by broadening a sequence of queries to obtain a query progression for each search. After the number of documents retrieved had been determined (i.e., retrieval size) for each query and the particular relevant document(s) retrieved from the recall base, an adjustment formula was applied. This resulted in the estimation of the number of documents that must be scanned before locating a particular document. By using the adjustment formula this number may be estimated from the gross retrieval size and recall figures. Details of this procedure are shown in Appendix I. A summary of the results is shown on the following page.

Recall level[2]

Effectiveness categories	25		50		75		100	
	D	FT	D	FT	D	FT	D	FT
Average retrieval size	32	38	72	86	140	157	252	296
Average precision (%)	41	34	36	30	28	25	21	18
Average fallout (%)	.0052	.0068	.0126	.0164	.0277	.0323	.0548	.0668

The two systems were then compared graphically as shown in Figure 1. It becomes even more clear that the Full-Text System involved consistently greater numbers of documents retrieved at any particular recall level than did the Descriptor System. These increases were not of sufficient magnitude to be considered statistically significant. The increase in numbers of documents retrieved was very slight (i.e., six documents), at low recall, moderate at middle ranges although higher at the highest recall level.

The broad range of results within each recall level is apparent from the individual retrieval figures given in Table IV. For instance, the Descriptor System retrievals at 25 percent recall ranged from 0 to 263 and those at the same level for the Full-Text System ranged from 0 to 144. This wide range pattern appeared to be consistent for both systems.

Another method of analysis concerning the individual results involved the cumulative distribution of retrievals within each recall level. Figures 2, 3, 4 and 5 display graphically the cumulative distributions within each recall level for the two systems. Although there was wide variation within each level it appeared that both systems were indeed similar at each level of recall. There was no wide variation between the two systems with regard to cumulative distribution at any point. This tended to reinforce the view that the two systems were comparable with regard to performance.

[2] Where D = Descriptor System
and FT = Full-Text System

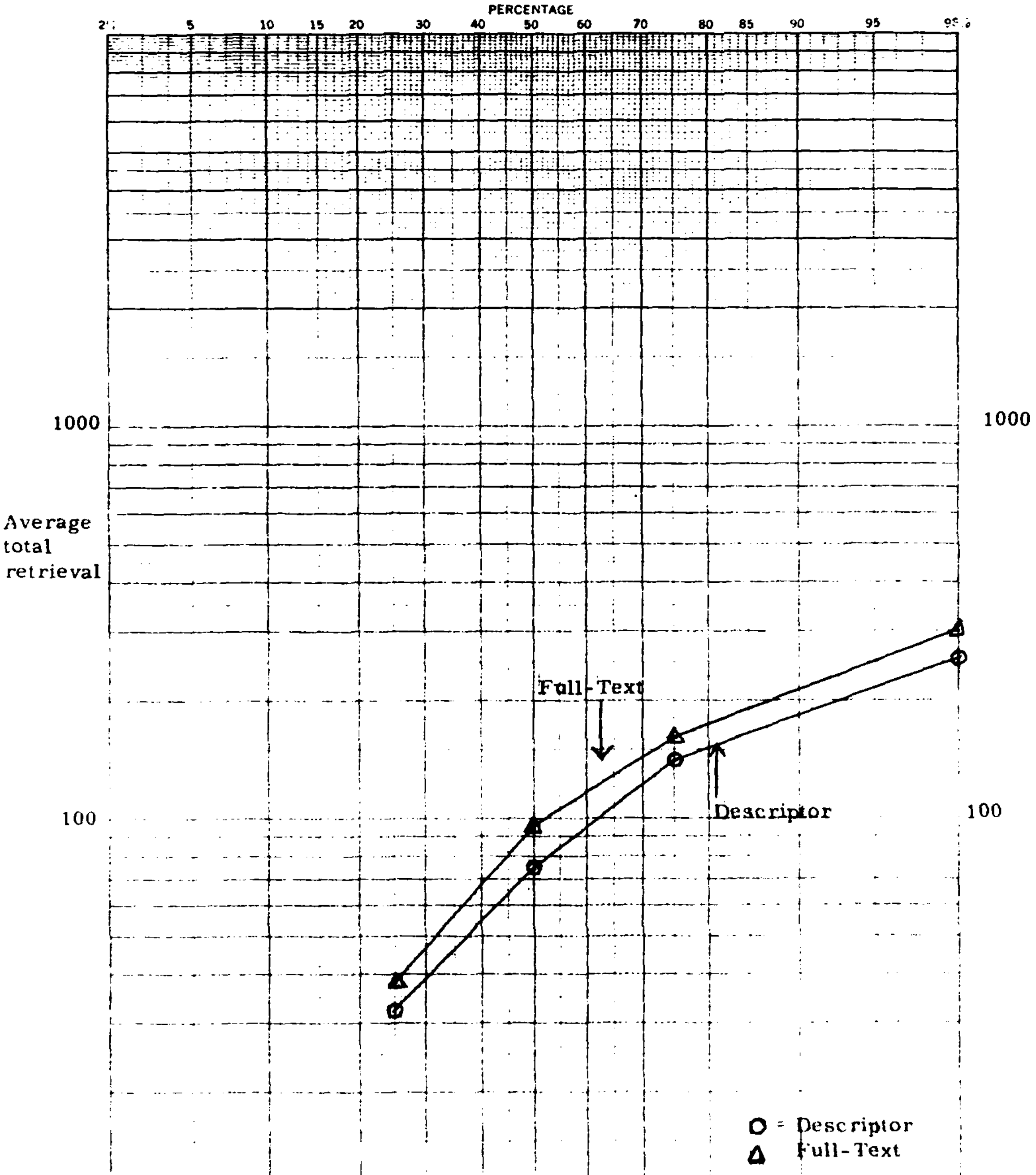

Figure 1 Comparison of documents retrieved by Full-Text and Descriptor Systems

Table IV Summary of adjusted retrieval* for the Descriptor and Full-Text Systems

Search	Recall percentage							
	25		50		75		100	
	D	FT	D	FT	D	FT	D	FT
1	2	27	38	76	72	100	286	462
2	263	144	650	384	777	483	1,003	592
3	0	0	0	0	0	0	0	0
4	0	1	1	2	2	2	4	16
5	126	129	253	172	1,038	652	1,697	1,440
6	2	1	3	3	54	4	305	6
7	76	91	165	181	179	577	250	969
8	3	1	6	3	146	4	285	6
9	7	6	21	13	58	45	97	116
10	5	7	10	14	14	22	19	29
11	1	5	2	11	2	23	57	41
12	6	4	11	7	17	11	22	14
13	38	38	81	77	100	144	146	221
14	11	16	21	40	101	49	171	70
15	2	3	5	9	17	43	128	213
16	4	12	9	24	13	179	53	632
17	28	44	57	87	72	89	518	91
18	7	10	17	21	20	31	74	40
19	87	122	173	339	260	437	346	883
20	8	9	15	41	23	77	30	104
21	13	85	26	187	105	206	194	285
22	20	8	40	15	60	117	80	317
23	4	35	12	108	18	182	22	254
24	60	119	119	250	220	280	262	298
Retrieval totals	773	917	1,735	2,064	3,368	3,757	6,049	7,099
Average retrieval	32	38	72	86	140	157	252	296
Standard error	12.0	11.8	3.5	5.6	17.3	14.7	18.3	22.4
Precision (%)	41	34	36	30	28	25	21	18
Fallout (%)	.0052	.0068	.0126	.0164	.0277	.0323	.0548	.0669

* where D = Descriptor System
and FT = Full-Text System

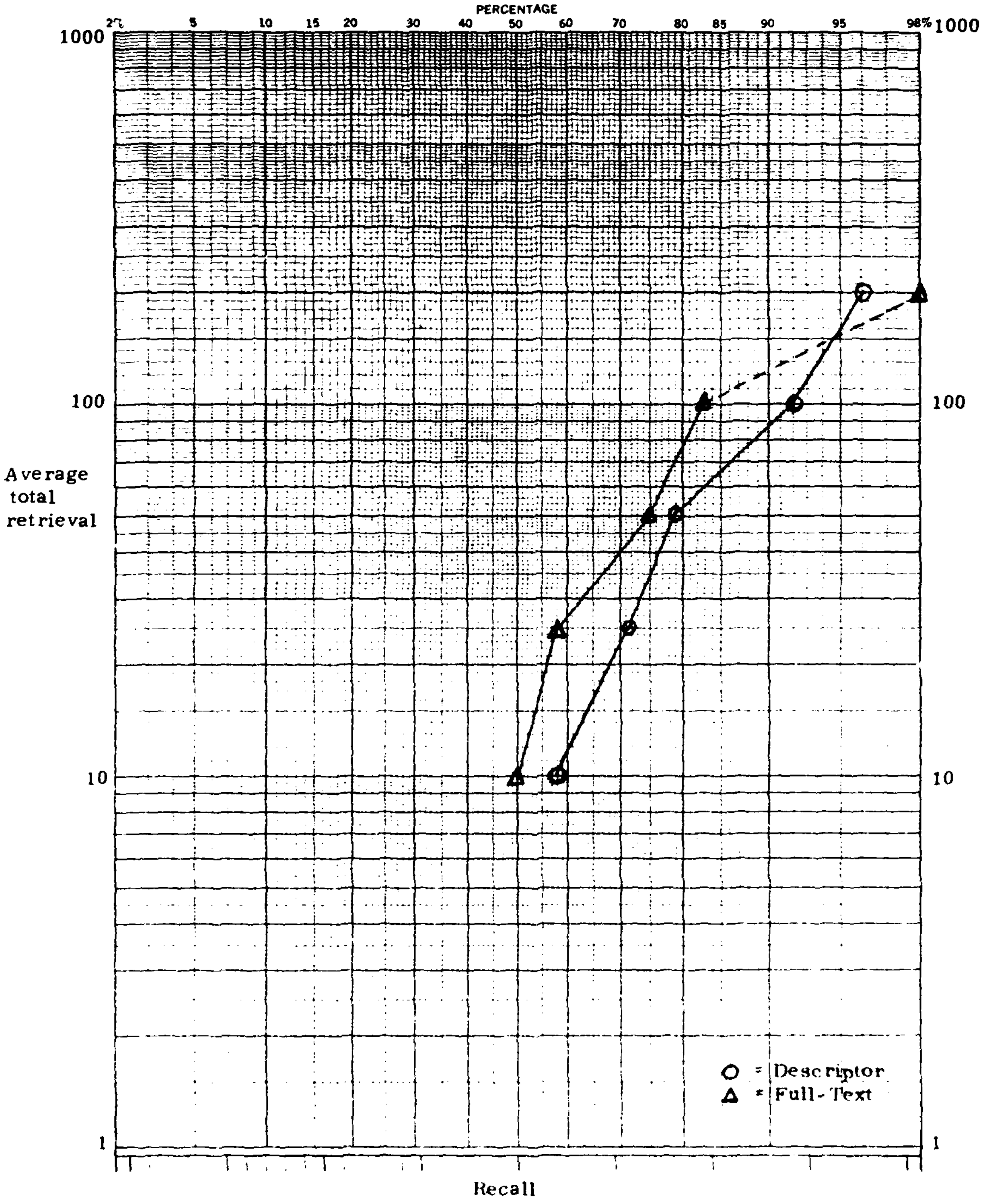

Figure 2 Cumulative distributions of Full-Text and Descriptor searches at 25 percent recall

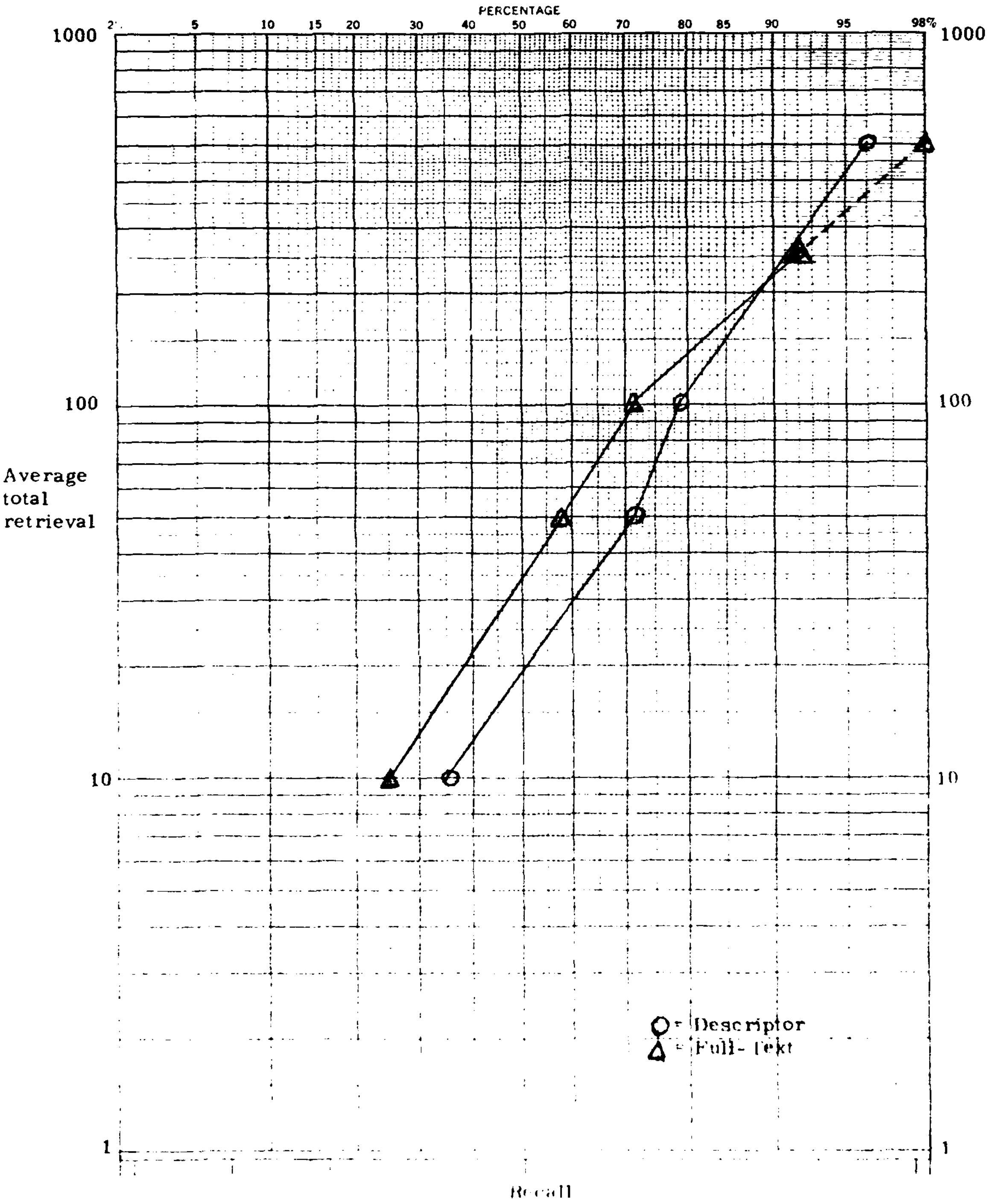

Figure 3 Cumulative distributions of Full-Text and Descriptor searches at 50 percent recall

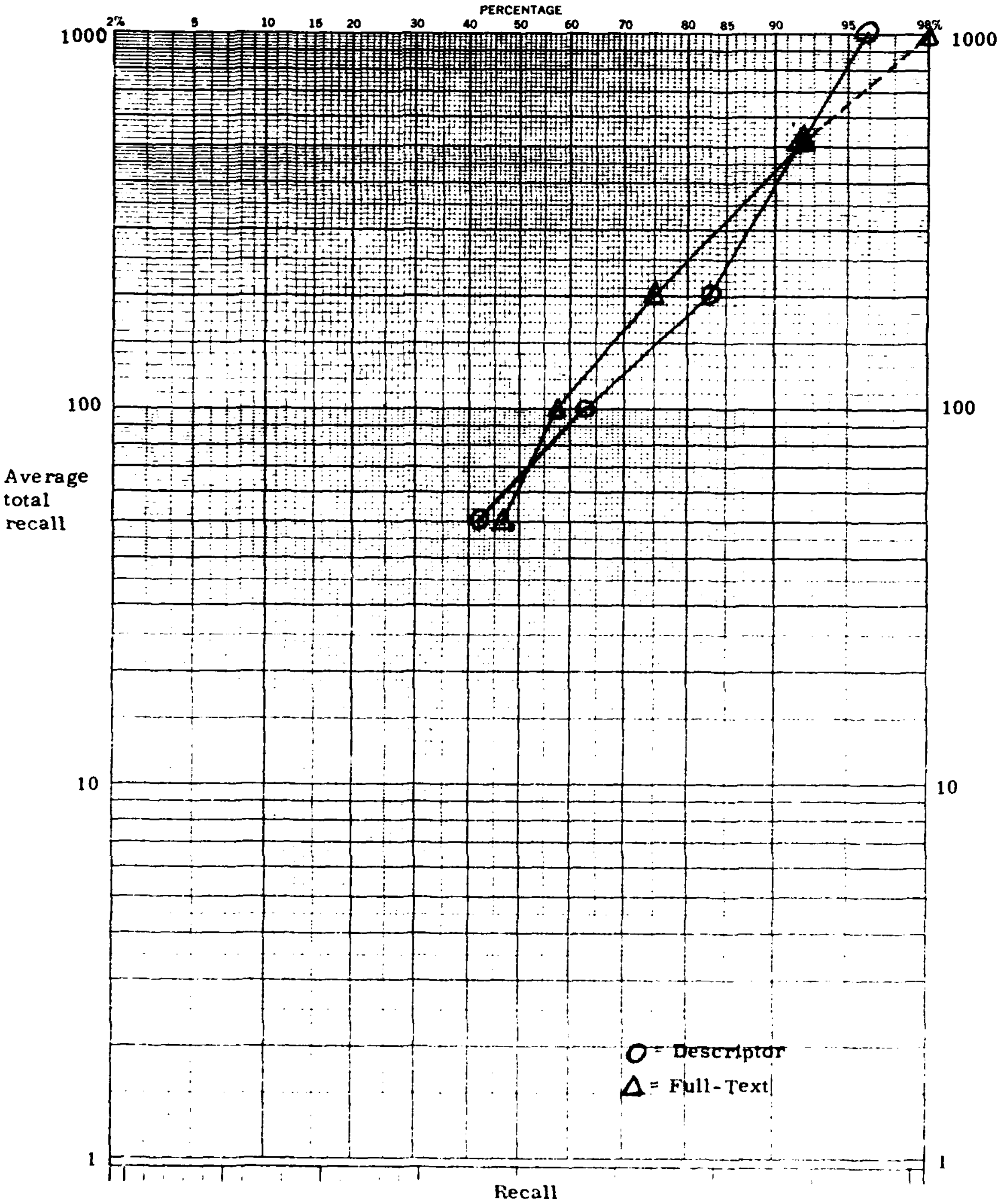

Figure 4 Cumulative distributions of Full-Text and Descriptor searches at 75 percent recall

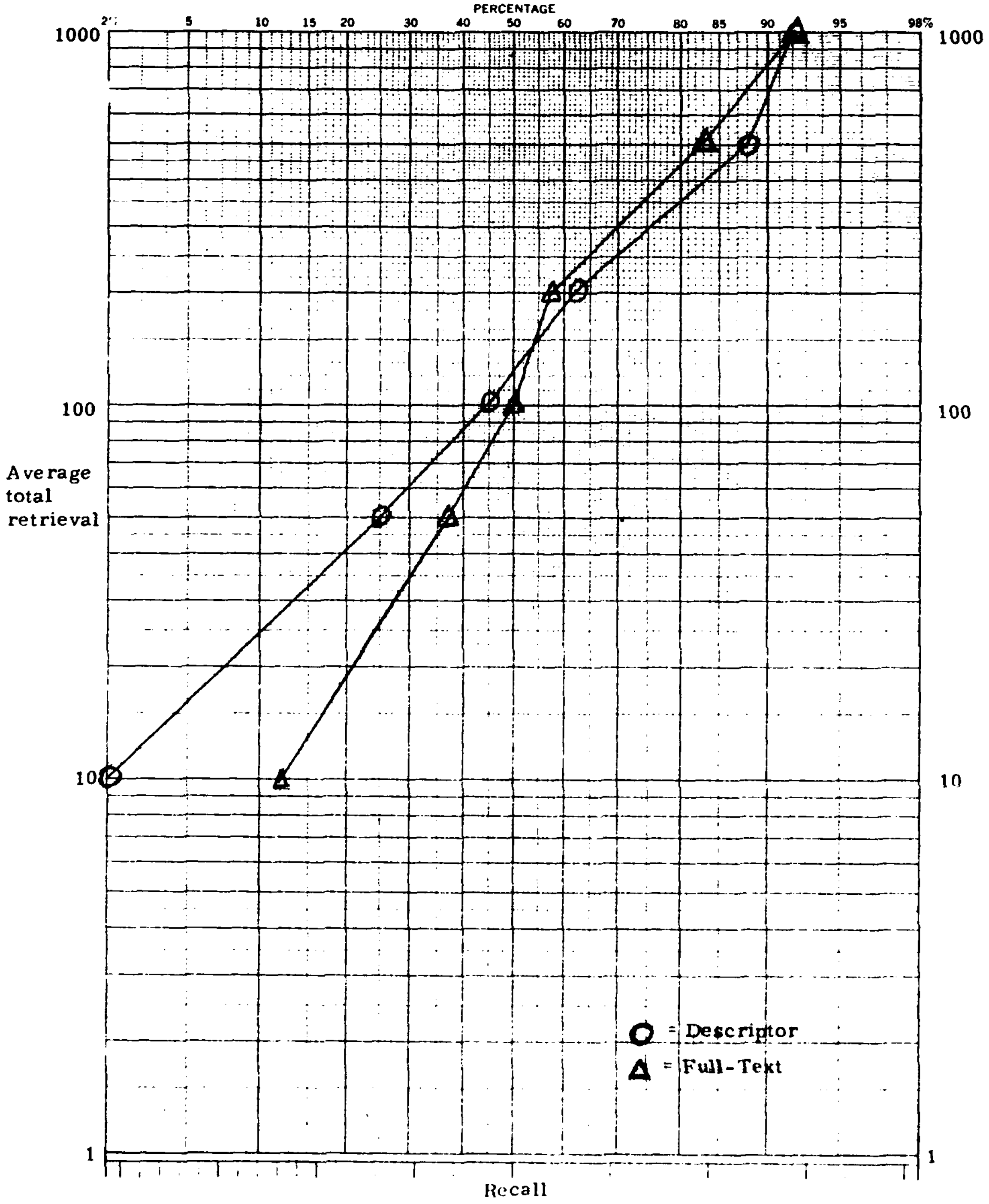

Figure 5 Cumulative distributions of Full-Text and Descriptor searches at 100 percent recall

Table V Gross retrieval results for 24 search queries for Phase 2, Part A

	Descriptor System						Full-Text System			
Search	Phase 1 query	Phase 2 query				Phase 1 query	Phase 2 query			
		25	50	75	100		25	50	75	100
1	3	6	97	187	172	11	42	69	380	300
2	21	503	378	57	144	286	9	22	294	25
3	4	4	0	0	0	0	2	80	0	0
4	4	0	7	11	10	2	0	0	9	18
5	20	353	358	495	1,625	257	5	260	788	790
6	0	94	4	99	403	6	14	494	383	383
7	6	150	37	55	57	578	0	271	32	156
8	8	0	2	274	3	3	0	3	9	0
9	15	13	29	30	19	52	0	18	45	45
10	23	0	1	19	44	71	0	0	0	35
11	55	0	0	2	53	15	162	15	15	20
12	21	0	2	9	4	17	3	9	11	14
13	20	74	13	38	73	621	0	114	59	95
14	14	3	17	173	209	31	9	276	26	25
15	29	3	3	20	174	9	5	6	63	268
16	13	3	2	3	72	82	55	35	268	621
17	55	1	1	29	862	85	2	0	2	3
18	26	51	1	21	22	40	0	0	0	71
19	4	765	6	423	428	242	1,178	291	9	651
20	3	11	28	34	34	46	17	46	25	28
21	38	90	6	42	40	276	169	54	60	62
22	99	19	10	20	116	188	2	22	188	211
23	10	6	5	50	7	240	69	77	71	72
24	178	1	3	13	125	49	236	26	53	75

A similar analysis was made of the gross retrieval figures as shown in Table V although the patterns were not as well defined and were therefore not as visible.

As was the case with the category of retrieval, precision was consistently slightly higher under the Descriptor System and fallout was consistently lower. However, these differences between the Descriptor and Full-Text Systems were not great enough to be statistically significant either.

Thus, after investigation of the performance of the Descriptor System and the Full-Text System in terms of recall, total retrieval, precision, and fallout, it was concluded that while there were differences between these two systems, these differences were extremely small and could not, therefore, be considered significant. For all practical purposes the two systems appeared to perform in a comparable manner and at a comparable level.

b. Part B - Stochastic Cost/Effectiveness Model

In order to evaluate various alternative retrieval systems it may be beneficial to subdivide the cost/effectiveness of these systems into their basic components that will allow manipulation of certain desired components in a variety of different combinations. In this model the three categories of components are as follows:

- User/system interface
- Query/system response
- Screening

There are several possible alternatives within each of these categories. However, for purposes of this evaluation a limited number are discussed.

The first category (user/system interface) included either the use or nonuse of an intermediary or search analyst. When an intermediary was used the interface could be through either written or oral contact.

The second category (query/system response) concerned, first of all, the choice of either a descriptor or full-text system. Within each of these there are four levels of operation (i.e., 25 percent, 50 percent, 75 per cent, 100 percent recall) and four associated fallout ratios under each system.

The third category (screening) included the choice of screening or no screening. Under the screening alternative comes the choice of screening representation to be used such as title and topic tags, or title and abstract and either loose or tight screening on each. Loose screening refers to a tendency on the part of the screener to pass nonrelevant information in an attempt to prevent the withholding of possible relevant information. In cases where there was doubt as to the nonrelevance of an item it would be allowed to pass the screener. Tight screening refers to the opposite strategy. In cases where doubt exists as to the nonrelevance of an item the screener would not pass the document. Tight screening occurs where there is a desire to reduce the number of documents to be scanned by the user even at the cost of withholding some relevant information.

Table VI shows the effectiveness probabilities that formed alternatives to be used in the model. For a detailed discussion of the model and its use refer to Section IV. Also in Section IV is an associated cost model[3] which is included for FTD use. The sample cost figures are partly hypothetical and are intended for demonstration only.

Three types of intermediary between intelligence analyst and query formulation were considered: (1) no intermediary, in which case the analyst wrote the query himself, (2) telephoned request to an intermediary, and (3) written request to an intermediary. The four levels of recall, 25, 50, 75 and 100 percent were considered and five types of screening: (1) loose screening on abstract and title, (2) tight screening on abstract and title, (3) loose screening on title and topic tags, (4) tight screening on title and topic tags and, (5) no screening. The results gave a total of 120 combinations for each system.

A few sample results are shown in Table VII.

Using the same alternative combinations, and combining both Descriptor and Full-Text Systems to create a hypothetical system more nearly comparable to CIRCOL at present, we find that recall remains the same while retrieval increases dramatically. For other alternative combinations the situation may change.

c. Part C - Recall Failure Analysis

In evaluating and comparing information retrieval systems two types of data must be analyzed. The first of these involves performance figures and the second search failures. It is this latter group that is considered in this section.

In general there are two types of search failures: (1) recall failures and (2) precision failures. Only recall failures are analyzed at this point

[3] Donald W. King and Nancy W. Caldwell, Cost Effectiveness of Retrospective Search Systems, American Phychological Association, March 1971.

Table VI Summary of effectiveness for various system alternatives

			$P(C_r/V_r)$[a]	$P(C_r/V_{\bar{r}})$[a]
Intermediary		1. none	.985	.0000011
		2. oral contact	.975	.0000017
		3. written contact	.950	.0000034
Query/System Response			$P(R_r/C_r)$[b]	$P(R_r/C_{\bar{r}})$[b]
	Descriptor	level 1	.25	.000052
		level 2	.50	.000126
		level 3	.75	.000277
		level 4	1.00	.000548
	Full-Text	level 1	.25	.000068
		level 2	.50	.000164
		level 3	.75	.000323
		level 4	1.00	.000669
Screening			$P(S_r/V_r)$	$P(S_r/V_{\bar{r}})$
	Titles and abstracts	1. loose	.8127[b]	.2871[b]
		2. tight	.5405[c]	.0051[c]
	Titles and topic tags	3. loose	.3495[d]	.1858[d]
		4. tight	.2322[c]	.0033[c]
	none	5. descriptor	1.000	1.000
		full-text	1.000	1.000

Where:

X_1 = 365,000 documents[b]

X_2 = 50,000 searches per year[b]

X_3 = number of items retrieved per search

X_4 = number of items mailed per search

X_5 = 2,200 terms in authority list

[a] Donald W. King and Nancy W. Caldwell, Cost Effectiveness of Retrospective Search Systems, American Psychological Association, March 1971.

[b] CIRCOL study

[c] P. Atherton, unpublished report on evaluation of document representations

[d] estimated by combination of notes b and c

Table VII Sample system alternative combinations using effectiveness model

Alternative combination	Number relevant documents retrieved	Number non-relevant documents retrieved	Recall (%)	Fallout (%)	Precision (%)
Descriptor System					
125*	25	47	50	.0126	35
225	25	46	50	.0126	35
325	25	47	50	.0126	35
Full-Text System					
125	26	60	50	.0164	30
225	26	60	50	.0164	30
325	26	60	50	.0164	30
Both systems combined					
125	26	124	50	.034	17
225	26	124	50	.034	17
325	26	124	50	.034	17

*The first digit corresponds to intermediary (1-none, 2-oral, 3-written); the second digit indicates Query/system response (recall level .25, .50, .75, 1.00); the third digit indicates screening mode (1-TA loose, 2-TA tight, 3-TT loose, 4-TT tight, 5-none).

due to the nature and availability of the necessary data as well as the relative importance of the two factors.

Although the causes of recall failures (nonretrieval of relevant documents) will vary from system to system and time to time, it may prove helpful in comparing two systems to categorize these recall failures and analyze possible consequences and methods of improvement.

The various sources of recall error are not distinct but are overlapping and highly confounded. Therefore, it is necessary to utilize fairly general categories in order to conduct any reasonable analysis.

For each failure it is necessary to examine the following material:

- the full text or abstract of the documents designated as relevant by the technical analyst prior to the actual search
- the assigned index terms for this document
- the request statement or technical analyst's statement of need
- the search formulation upon which the search was conducted

In evaluating recall failures of the Descriptor (Index) System and the Full-Text Natural Language System, search queries were formulated under both systems for 30 requests as part of Phase 1. At that time recall (proportion of relevant documents retrieved) estimates were calculated for each query through the use of a recall base designated by the request originator. For each query the portion of the appropriate recall base that was not retrieved by that query is known. Also which documents make up that portion of the recall base is known. Each query then may be examined in association with its unretrieved portion of the recall base in order to determine the reason(s) for failure to retrieve the relevant documents. This examination is carried out separately for each query, independently for each of the two systems.

In assigning individual recall failures to the various categories, an attempt was made to determine the basic underlying cause of the particular error. Some instances were encountered that involved more than one source of failure. For a small proportion of these cases it would have been impractical to attribute the failure to only one of two failure categories; therefore, the failure was divided equally between the two categories with one-half of the failure attributed to each.[4]

(1) Descriptor System failure categories

The five broad categories of recall failure that were used for the Descriptor System involved failures revolving around the:

- Index Language
- Indexing Process
- Searching Process
- User/System Interaction, or
- Other

The first category (Index Language) encompasses such factors as:

- Inadequacy of hierarchy
- Lack of specific terms, and
- Lack of appropriate word endings (WFC File)

The second category (Indexing Process) involves:

- Lack of specificity
- Lack of exhaustivity
- Omission of important concepts, and
- Use of inappropriate terms

Category three (Searching Process) includes:

- Failure to cover all reasonable approaches
- Formulation too exhaustive, and
- Formulation too specific

The fourth group (User/System Interaction) concerns only that portion of the written request statement which relates to the interface between user and system intermediary or search analyst. Failures included

[4] F. W. Lancaster, Information Retrieval Systems, John Wiley & Sons, 1968. p. 134

in this group indicate that the user's request statement was in some way different from his actual need. In other words, the request statement was either too broad, too narrow, or imperfect in some other respect.

Other failures arose primarily from clerical or keyboarding errors.

(2) Full-Text Natural Language System failure categories

Four broad categories of recall failure used in evaluating the Full-Text Language System involved:

- Synonym Failures
- Searching Process
- User/System Interaction, and
- Other

The first category (Synonym Failures) is roughly analagous to the problems associated with indexing under a descriptor system and includes:

- Failure to retrieve a related or similar term as well as usual synonyms
- Failure to retrieve various term endings (WFC File)

These failures are attributable to the Full-Text System. The searcher did not cover all possible approaches (i.e., consider all synnyms or word forms); however, it is felt that more beneficial analysis may be gained from treating this as a separate category rather than including these failures under the Searching Process category of "failure to cover all reasonable approaches".

The second category (Searching Process) is composed of the same three failures as under the Descriptor System:

- Failure to cover all reasonable approaches
- Formulation too exhaustive, or
- Formulation too specific

The third category (User/System Interaction) concerns, as previously, the written request which formed the interface between user and system intermediary. Failures in this category indicate an imperfect request statement.

Other failures concern clerical or keyboarding errors.

Since there may be some degree of ambiguity associated with the terms used to describe the various categories of recall failure it should prove beneficial for total understanding of the analysis to give examples of specific failures under each category. Following is a brief list of examples taken from the CIRCOL System:

Index Language

- Endings (WFC file) - lack of such terms as "sustaining" although the dictionary includes "sustained" and "sustainer"
- Lack of specific terms - no term for "omegatron"
- Inadequate hierarchy - lack of further breakdown of such categories as "VSTOL aircraft" into such narrower terms as "G222", "Fiat", or "G91y".

Indexing Process

- Lack of specificity - document indexed under the term "jet engine" when document refers to "air-jet-engines" or an article on the Soviet coordinate computing center is indexed under "space coordinate tracking".
- Lack of exhaustivity - since only a limited number of terms are chosen for each document, a document discussing various alloys and mentioning the use of forgings is indexed only under the various types of alloys (i.e., aluminum base alloy) and not under "forging".
- Omission of important concepts - a document referring to Soviet space stations is indexed under only "space tracking" and "computer center".
- Use of inappropriate terms - a document concerning the animal "badger" is indexed under "badger aircraft".

Searching Process

- Failure to cover all reasonable approaches - a search request for information about rocket launching in the Mediterranean is searched by "rocket launching" and "Mediterranean" and not under all the individual surrounding countries, such as Egypt, Turkey, etc. Or, a request for physical changes during spaceflight is searched on "physiological parameter" and "spaceflight" and not such terms as "body weight", "metabolism", or "heart rate".

- Formulation too exhaustive - a request asking for cooling off components in propulsion systems is searched by "rocket engine" and "thermodynamic" and "cooling" (in an 'and' relationship).

- Formulation too specific - the same request above for cooling of components in propulsion systems is seaṙched by "cooling" and "propulsion system" and "component".

User/System Interaction

- Request different from actual information need - a request for metal removal lists relevant documents on hole drilling in metals, or, a request for information on Soviet experiments in the area of space life support systems in 1968 lists relevant documents from 1966 and 1967.

Other

- Clerical and keyboarding errors such as misspelling of words and typographical mistakes as "Feburary" or "conveneince".

Synonym Failures

- Failure to consider a related term - the search term "corporation" will not retrieve the term "company".

- Failure to consider term endings - the search term "analyze" will not retrieve the term "analyzer".

Table VIII summarizes the results of the recall failure analysis for the 30 requests under the Descriptor System and Table IX under the Full-Text Natural Language System.

Table X displays summary results for both systems.

Table VIII Number and percentage of total recall failures under the Descriptor System

Category	Failure Type	Number of Occurrences*	Percent of total failures
Index Language	Lack of word endings (WFC file)	6	7
	Lack of specific terms	1	
	Inadequate hierarchy	0	
Indexing Process	Lack of specificity	2	20. 5
	Lack of exhaustivity	7. 5	
	Omission of important concepts	11	
	Use of inappropriate terms	0	
Searching Process	Failure to cover all reasonable approaches	20. 5	49
	Formulation too exhaustive	5	
	Formulation too specific	23. 5	
User/System Interaction	Request different from actual information need	23. 5	23. 5
Other	Clerical or keyboarding	0	0
Total			100. 0

* The number of occurrences is different for the two systems due to the fact that a different number of relevant documents was "missed" by each search query, thus yielding a different number of failures under the two systems.

Table IX Number and percentage of total recall failures under the Full-Text Natural Language System

Category	Failure type	Number of occurrences	Percent of total failures
Synonym Failures	Failure to consider related, similar terms or synonyms	7	12
	Failure to consider a term ending (WFC file)	0	
Searching Process	Failure to cover all reasonable approaches	12	
	Formulation too exhaustive	0	38
	Formulation too specific	11	
User/System Interaction	Request different from actual information need	30	50
Other	Clerical or keyboarding	0	0
Total			100

Table X The percentage of errors for both systems shown together for comparison purposes

System	Percent of recall failures					
	Index Language	Index Process	Synonym Failures	Searching Process	User/System Interaction	Other
Descriptor	7	20.5	---	49	23.5	---
Full-Text Natural Language	---	---	12	38	50	---

Under the Descriptor System, the prime category of failure appears to be the Searching Process with a tossup between Indexing Process and User/System Interaction for the second most frequent cause. Close to one-half of the recall failures under the Descriptor System fall into the category of the Searching Process. Most of these were due to either the formulation being too specific or failing to cover all reasonable approaches. Relatively few of these errors were caused by excessive exhaustivity. Under Indexing, most of the recall failures were due to the omission of important concepts or to a lack of exhaustivity. Failures caused by User/System Interaction indicate that the written requests differed (usually in the level of specificity) from the user's actual information need.

Under the Full-Text Natural Language System, the prime cause of recall failure was User/System Interaction. In other words, in approximately 50 percent of the cases the request simply was not an accurate statement of the real information need. This was due to either the level of specificity or generality, the lack of necessary descriptive details, or the inclusion of unnecessary superfluous description. The second most frequent category of failure under the Full-Text System was the Searching Process. These failures were due to the formulation being too specific or failing to cover all reasonable approaches. This was the same under the Descriptor System. Neither the Synonym Failure nor Other categories appear to be significant.

It is meaningful that under both systems, Descriptor and Full-Text, the two principal categories of failure are the Searching Process and the User/System Interaction.

Looking in detail at these two categories under the Descriptor System, it is possible that the Searching Process was a major cause of recall failure due to the difficulty involved in transposing a written information request into the proper controlled vocabulary terms. Not only must the appropriate concepts be chosen, but the proper hierarchical level must be pinpointed as well. Often in an attempt to retrieve all possible relevant documents while at the same time obtaining a feasible number of documents retrieved, the search query may be narrowed of necessity. This achieves the goal of having an acceptable number of documents retrieved but also reduces the number of relevant documents retrieved. Thus, in making the proper tradeoff decision, recall may become the element which suffers.

User/System Interaction as a source of failure is closely related to the problems incurred in the Searching Process. The written request may, in many cases, fail to reflect the actual user's information need. The request may indicate that a certain level of specificity is desired when actually the user is looking for another. Other cases occur in which terms used in the request may be either misleading or misinterpreted. Many times it is extremely difficult in subsequent analysis to determine whether the fault rests with too specific a search query or too general a request statement. No matter where the responsibility for failure is placed it is obvious that a principal cause of recall failure concerns the communication process occurring between user and system.

With regard to these two categories (User/System Interaction and the Searching Process) under the Full-Text Natural Language System, some of the causes for failure might be expected. Failures caused by User/System Interaction are approximately 50 percent of the Full-Text System failures. These could be due to the fact that when writing the request statement the users are mentally geared to using the Descriptor System, and, therefore, unconsciously choose terms that could be appropriate under that system but that are not reflective of their true information need when using a full-text natural language search query. Users may be accustomed to expressing their need in either more general or more specific terms than actually necessary in order to adjust to the peculiarities of the present operating system (Descriptor).

Under the Searching Process, which is the next most frequent category of failure (38 percent) within the Full-Text Natural Language System, the sources of the most often occurring failures appear to arise because the formulation was too specific or failed to cover all reasonable approaches.

As with the Descriptor System these two categories are closely related to the communication process involving the user and system and it is difficult to place the initial responsibility for failure precisely in one of those two categories.

It is interesting to note that while the same two categories are involved as the major sources of failure under both the Descriptor and Full-Text Natural Language Systems, the order of occurrence is reversed. Under the Descriptor System the Searching Process failures occur more frequently than those involving the User/System Interface while under the Full-Text System the User/System Interaction is the source of more failures than the Searching Process. This indicates that improvement in recall may be easier under the Full-Text System than under the Descriptor System. Many avenues are open for improvement in, and alteration of the process of User/System Interaction which may be brought about with relatively little expenditure of time or funds. For instance, new, or even old, system users could be briefed on the possible pitfalls concerning the formulation of search requests. Training could be offered to aid in determining the proper method of ascertaining the most efficient hierarchical level to pursue and to develop procedures for further modification of the search strategy depending on the nature of the initial search strategy results.

Failures in the Descriptor System differ in their indication from those under the Full-Text System in that the user's request statement is less at fault than the search query formulation. This may mean that the user experiences considerable difficulty in attempting to formulate a Descriptor Search even when the true information need is understood. While there are various methods of obtaining improvement these may prove to be time consuming and costly. For instance, the major source of improvement may lie in reformulation of the index language or retaining of the index personnel, etc.

Thus, while the sources of recall failures are by no means clear-cut or well defined, certain trends may definitely be observed and measures taken to improve the recall performance of the system.

d. Part D - Effects of Analyst Experience

(1) Purpose and Procedure

An experiment was designed to isolate the effect on search performance of two factors. The first of these factors was the extent of CIRCOL experience of the analyst and the second was the use of relevant documents during the formulation of the query. Performance was measured in terms of recall (proportion of relevant documents retrieved) and total retrieval. Performance is assumed to increase as recall is maximized and retrieval is minimized. Thus, the "best" search query would be that which exhibits the highest possible recall ratio and the lowest possible number of documents retrieved.

CIRCOL personnel selected four analysts to participate in the experiment. Two were experienced in the use of CIRCOL and two had little or no experience with CIRCOL. These four met as a group and agreed upon five fairly narrow search topics. They also jointly identified six documents which were relevant to each search topic.

Each analyst, working independently of the others, formulated a search query for each of the five topics. When these queries were completed each analyst was provided with a common pair of relevant documents out of the originally named six. These documents were used by the analyst to formulate a second search for each topic. After seeing the two relevant documents, the analyst had the option of not writing a second query if he felt that his original query was still the best he could do.

In all cases analysts were advised to use natural language terms rather than topic tags and to formulate their queries using terms likely to occur in actual abstracts rather than topic tags. Recall was estimated on the basis of terms which appeared in the actual text (and title) of each document -- not on topic tags.

The resulting queries were run on CIRCOL by Westat. The total number of documents retrieved by each query was recorded. Each query was also matched against four of the originally specified relevant documents for that search in order to determine recall. The four documents used in determining the recall ratio did not include the two which were used by the analysts in reformulation of queries.

(2) Results

Results of the experiment are shown in Table XI. Zero retrieval with up to 100 percent recall was possible because not all of the relevant documents named by the analysts were actually in the CIRCOL database. Individual review of each document determined whether it would have been retrieved by a query had it been in the CIRCOL database.

For the initial queries, inexperienced analysts averaged 28 documents retrieved with recall at 35 percent. Experienced analysts retrieved an average of 83 documents with recall at 48 percent.

When second queries are considered, the inexperienced analysts slightly increased recall to 40 percent, and also increased retrieval to 37 documents.

Table XI Recall and retrieval results for analysis of analyst experience

Query	Percent Recall				Retrieval			
	Experienced analysts		Inexperienced analysts		Experienced analysts		Inexperienced analysts	
	#1	#2	#3	#4	#1	#2	#3	#4
Topic 1								
1st query unmodified	100	25	75	100	190	11	7	54
2nd query* modified	100	75	50	50	190**	8	7	0
Topic 2								
1st query unmodified	100	50	25	50	68	8	5	0
2nd query* modified	100	100	25	100	68**	14	5**	0
Topic 3								
1st query unmodified	0	0	0	25	58	55	73	0
2nd query* modified	0	0	0	25	58**	18	73**	3
Topic 4								
1st query unmodified	75	50	75	0	95	28	129	3
2nd query* modified	100	50	75	0	130	28**	129**	2
Topic 5								
1st query unmodified	75	0	0	0	277	37	4	1
2nd query* modified	75	0	0	75	277**	37**	4**	145
Average for								
1st query	70	25	35	35	138	28	44	12
2nd query	75	45	30	50	145	21	44	30
Average for experience level								
1st query unmodified	48		35		83		28	
2nd query* modified	60		40		83		37	

* modified during search process using pair of relevant documents

** analyst used the same query and chose not to modify

Experienced analysts had better success in approaching a "best" search. Average retrieval was held constant but recall increased from 48 percent to 60 percent.

Queries were actually modified only 50 percent of the time. It is assumed that after consulting the two relevant documents, the analysts still felt that the other half of the time their original query could not be improved upon. Experienced analysts modified queries somewhat less frequently (40 percent of the time) than inexperienced analysts (60 percent of the time).

(3) Analysis

This experiment shows that, in general, recall tends to be improved when relevant documents are used as an aid to query formulation. This improvement was more noticeable for experienced analysts than for inexperienced ones. However, there were some specific cases in which recall actually decreased when the query was modified after seeing relevant documents. This was only true with inexperienced analysts, which could account for the smaller improvement for inexperienced analysts than for experienced ones.

(4) Whether or Not to Modify

It might be assumed that experienced analysts were less likely to modify their queries because their experience had taught them to write fairly effective queries initially. However, there are at least two other factors which might affect their decision to modify a search on the basis of seeing relevant documents.

The first factor would be a knowledge of number of documents retrieved by the original query and the recall ratio attained. Low recall and/or high retrieval might have resulted in a decision to modify the query. In cases where recall was zero and queries were not modified, it should be safe to assume that queries would have been modified if the analysts had known they were getting zero recall.

For those (10) queries which were not modified, five queries would not have retrieved either of the two relevant documents which the analysts had, two would have retrieved only one of the two relevant documents. Thus, some factors other than these two relevant documents were deciding factors in the decision not to modify. In only three cases (of these 10) were original queries actually retrieving both of the relevant documents.

For those (10) queries which were modified, five picked up the same number of relevant documents (of the two used by the analysts).

Three of these were already retrieving both relevant documents, while two retrieved neither document with the original query nor with the modified one. Both of these latter cases resulted from queries written by inexperienced analysts. In addition, in all three cases where original queries were retrieving both relevant documents, the recall ratio (when calculated on the other four relevant documents) was increased. Therefore, the use of these relevant documents as query formulation aids can be considered of some value. In the two cases where neither relevant document was retrieved by either the original or the modified query, recall remained constant or decreased.

Four of the original queries retrieved neither relevant document. When these two documents were used as aids to formulate a modified query, three queries retrieved one relevant document, while one (by an experienced analyst) retrieved both. One query (by an inexperienced analyst) retrieved both relevant documents in their original form, but when modified, retrieved neither.

Changes in recall rates, however, were not consistent with whether or not more or fewer of the "aiding" documents were retrieved. For example, recall improved in only two of the four cases where number of "aiding" relevant documents retrieved was increased.

Individual experience was generally erratic although, when averaged over all cases, use of relevant documents in the writing of queries was helpful in improving recall.

e. Part E - Analysis of the Effect of File Additions

(1) Description and Purpose

An analysis of two proposed file additions to CIRCOL was conducted to provide some insight into the possible effects that might be expected. The two files considered were a Word Form Conversion (WFC) File and a Synonym/Equivalent (S/E) File. The WFC File is proposed as a mandatory addition to CIRCOL, i.e., one which would not be optional for the user. The S/E File would be optional for the user.

To simulate the effects of using one or both files original queries were expanded to include additional terms. The results of conducting these expanded queries were examined for changes in total retrieval and recall ratio.

(2) Procedure

(a) Query Formulation

Original queries from Phase 1 were expanded for both a Descriptor System (the first Phase 1 System) and Full-Text System Natural Language (Need) of Phase 1. To produce queries which would simulate a system containing the WFC File, terms were added to the original query that were considered to be precise single-word synonyms (at all times), variant spellings, common misspellings, and a few specific endings for both verbs and nouns. These queries were then further expanded to produce queries which would simulate the additional use of a S/E[5] File. Terms added were acronyms, multiple-word synonyms including expanded acronyms, narrower terms and related terms.

The most severe practical problem encountered in formulating expanded queries was that a number of them become too lengthy; i.e., when all terms were added, the maximum number of permissible lines for CIRCOL was exceeded. For most of these it was possible to reduce the number of lines by running individual terms and eliminating those which retrieved zero documents. One search could not be reduced by this method. It was therefore eliminated from the sample for this portion of the study. A second search was eliminated from the sample when it was found to be impossible to obtain consistent retrieval from CIRCOL.[6]

(b) Processing Queries

Each of the remaining 132 queries (3 Full-Text queries and 3 Descriptor queries for each of 22 searches) was run on the entire CIRCOL database and the number of documents retrieved recorded. Each was also qualified by DPSNR GE 365000, and this retrieval recorded.

(c) Results

For Descriptor System queries the number of documents retrieved with DPSNR GE 365000 was subtracted from total retrieval, giving a number of documents retrieved on a 365, 000-document database.

To simulate a 365, 000-document database for a Full-Text System (in which all documents are assumed to contain abstracts) the same method was used as for Descriptor System, but the resulting

[5] S/E File use implies the concurrent use of the WFC File.

[6] See Appendix III.

retrieval number was then multiplied by a ratio of 1.54. This was the same method as used in Part A. A description and example may be found in the Appendices I and II.

Increases in number of documents retrieved were substantial, from an average on original searches of 29 documents for Descriptor System queries and 132 documents for Full-Text System queries. Retrieval with both files in operation averaged 48 documents and 274 documents respectively. Thus, it is clear that the file changes result in substantially larger retrievals. Individual retrieval results are shown in Table XII.

Recall was also estimated using the same method as in Part A. The same sample of four relevant documents for each search was used. Each document was compared with the six individual queries (Descriptor-Original, WFC, WFC and S/E; Full-Text-Original, WFC, WFC and S/E) for that search and the number which would have been retrieved recorded.

Out of a total of 44 expanded queries increases in recall were experienced in only four cases. Two of these occurred under the Full-Text System, one from WFC and one from S/E File additions; two occurred in the Descriptor System, and again one each from each file addition.

Increases in recall involved a single term in each case. The increases using the WFC File were for one synonym from Roget's Thesaurus (in a Full-Text query) and one synonym from the CIRC Thesaurus (in a Descriptor query). S/E increases were a related term (in a Full-Text query) and a narrower term (in a Descriptor query) both from the CIRC Thesaurus. Individual recall results are shown in Table XIII.

Table XII Retrieval results for simulation of file changes in Phase 2 Part E based on 365, 000 document database

Phase 2 Search	Descriptor			Full-Text		
	Original query	With WFC	With S/E	Original query	With WFC	With S/E
1	---	---	---	---	---	---
2	---	---	---	---	---	---
3	4	4	4	0	0	0
4	4	14	28	2	2	2
5	20	74	74	257	257	277
6	0	0	0	6	6	6
7	6	6	6	578	579	581
8	8	20	47	3	6	9
9	15	16	16	52	75	75
10	23	23	43	71	71	74
11	55	56	57	15	25	25
12	21	22	22	17	17	17
13	20	20	19	621	634	2,335
14	14	36	41	31	37	71
15	29	29	29	9	9	9
16	13	63	237	82	439	439
17	55	55	58	85	85	89
18	26	26	26	40	40	40
19	4	4	4	242	242	659
20	3	3	3	46	46	85
21	38	38	40	276	710	724
22	99	99	108	188	188	205
23	10	10	10	240	245	254
24	178	178	178	49	60	62
Total	645	796	1,050	2,910	3,773	6,038
Average	29	36	48	132	172	274

Table XIII Recall results for simulation of file changes in Phase 2 Part E based on 365, 000 document database

Phase 2 Search	Descriptor			Full-Text		
	Original query	With WFC	With S/E	Original query	With WFC	With S/E
1	---	---	---	---	---	---
2	---	---	---	---	---	---
3	1/4	1/4	1/4	4/4	4/4	4/4
4	4/4	4/4	4/4	3/4	3/4	3/4
5	0/4	0/4	0/4	2/4	2/4	2/4
6	0/4	0/4	0/4	4/4	4/4	4/4
7	0/4	0/4	0/4	2/4	2/4	2/4
8	2/4	2/4	2/4	4/4	4/4	4/4
9	2/4	2/4	2/4	3/4	3/4	3/4
10	4/4	4/4	4/4	4 4	4/4	4/4
11	0/4	0/4	0/4	2/4	2/4	2/4
12	4/4	4/4	4/4	4/4	4/4	4/4
13	1/4	1/4	1/4	2/4	2/4	3/4
14	0/4	1/4	1/4	1/4	1/4	1/4
15	3/4	3/4	3/4	2/4	2/4	2/4
16	0/4	0/4	0/4	2/4	3/4	3/4
17	1/4	1/4	2/4	0/4	0/4	0/4
18	3/4	3/4	3/4	3/4	3/4	3/4
19	0/4	0/4	0/4	1/4	1/4	1/4
20	0/4	0/4	0/4	2/4	2/4	2/4
21	2/4	2/4	2/4	4/4	4/4	4/4
22	4/4	4/4	4/4	4/4	4/4	4/4
23	2/4	2/4	2/4	4/4	4/4	4/4
24	2/4	2/4	2/4	2/4	2/4	2/4
Total	35/88	36/88	37/88	59/88	60/88	61/88
Recall ratio (%)	40	41	42	67	68	69

(d) Analysis

Fairly large increases in number of documents retrieved were experienced. At the same time only minimal improvement in recall occurred. A summary of recall and retrieval increases is given in Tables XIV and XV. This was true for both the Descriptor System and for the Full-Text System. Discrepancies between the two increases were greatest under the Full-Text System where, with the addition of the WFC File, recall increased only 1 percent, while retrieval increased 30 percent. If both WFC and S/E Files were used, recall increased only 3 percent, but retrieval increased by 108 percent. On the average, for a small increase in proportion of relevant documents retrieved by the system, total retrieval doubled.

In terms of number of cases where increases actually occurred, the differences between increases in recall and retrieval were even more noticeable. Again considering only a Full-Text System which incorporates the WFC File, recall improved in only one case out of the sample of 22 cases, or 5 percent of the time. Increases in retrieval occurred in 10 cases, or 45 percent of the time. Thus, 9 cases, or 41 percent experienced an increase in number of documents retrieved without any concurrent increase in number of relevant documents retrieved. Considering the same system and having the user choose the option of the S/E File in addition to the WFC File, improvements in recall still occurred in 4 percent of searches, while retrieval increases occurred 59 percent of the time. The situation was not noticeably improved if both file additions were considered to be optional.

Table XIV Full-Text System occurrences of recall and retrieval increases

- Using WFC File.

  Increase in recall occurred one time, or 5 percent of the time
  Increase in retrieval occurred 10 times, or 45 percent of the time

- Using WFC File and S/E File.

  Increase in recall occurred two times, or 5 percent of the time
  Increase in retrieval occurred 23 times, or 52 percent of the time

- Both files, optional.

  Increase in recall occurred in two cases, or in 9 percent of cases
  Increase in retrieval occurred in 16 cases, or in 73 percent of cases

- Using S/E File (increases from WFC).

  Increase in recall occurred one time, or 5 percent of the time
  Increase in retrieval occurred 13 times, or 59 percent of the time

Table XV Summary of recall and retrieval increases for Phase 2, Part E file changes

Effectiveness categories	File	Number of documents retrieved (average)	Increase from original (average)	Increase from WFC (average)	Average percent increase from original	Average percent increase from WFC
Descriptor System retrieval	Original	29				
	WFC	36	7		24	
	S/E	48	19	12	66	33
Recall percent	Original	40				
	WFC	41	1		2	
	S/E	42	2	1	5	2
Full-Text System retrieval	Original	132				
	WFC	172	40		30	
	S/E	274	142	102	108	60
Recall (percent)	Original	67				
	WFC	68	1		1	
	S/E	69	2	1	3	3

SECTION IV

COST/EFFECTIVENESS MODEL

The purpose of this section is to set forth a cost/effectiveness framework for choosing among alternative information search and retrieval systems. It is hoped that this framework can be used by systems management to make better informed decisions concerning alternative systems and processes. System management is nearly always faced with these decisions in view of numerous alternatives available for a specific process as well as the highly interactive effect each of these may have on other processes in the overall system. A cost/effectiveness model has been derived to cope with the complex nature of information search and retrieval systems. This model permits one to investigate a broad range of alternate system combinations based on a minimum of primary experimental data. Thus, if system management is considering the feasibility of an entirely new system, a major portion of the alternate combinations can be discarded as unfeasible from a cost/effectiveness standpoint and the research effort can then be concentrated on those combinations that are likely to yield optimum cost/effective results. If system management observes an operating system, the model can be used to diagnose the system in order to pinpoint system components that can achieve better cost/effective results.

Basically, the model combines a number of information input, search, and output processes and provides cost and effectiveness parameters that are identified for each of the processes. Parameters related to cost include such factors as number of items input, number of searches, number of items retrieved per search, number of items sent to users, and number of terms in the authority list. Effectiveness is measured by probabilities that correspond to recall and fallout. The model then is used to estimate total retrieval and number of relevant items retrieved which presumably can be related to value in order to establish system benefits. Trivial relationships for cost/effectiveness are given as total cost, cost per search, and cost per relevant item retrieved. Several examples are given in which cost information is based on secondary sources and effectiveness measures for several subsystem processes are partially derived from the experiment described previously for comparative evaluation of the retrieval effectiveness of descriptor and free- text search systems using CIRCOL (Central Information Reference and Control On-Line) and from some secondary sources.

It is emphasized that the hypothetical costs used in the examples are derived from general sources and they do not reflect the costs associated with CIRCOL.

1. Document search and retrieval systems

Most document search and retrieval systems involve a triad of input, search, and output subsystems. The input subsystem usually consists of some combination of processes necessary to translate natural language of a full text into document representations (title, bibliographies, abstracts), index terms, computer codes and sometimes a computerized associative term file.

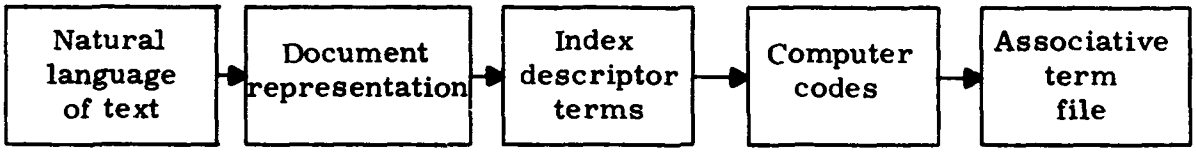

Figure 6 Stages of document processing in the input subsystem

The processes for each stage above include preparation of abstracts (or indexes), reproduction, keypunching, storage and so on. There are any number of combinations of stages of a system that may apply. For example, indexing may involve the full text of a document or indexing may only be from a document representation such as an abstract.

On the other hand, search subsystems often concern translation of a natural language statement of search need by intermediary interpretation, search query formulation in system terminology, computer codes and sometimes a computerized associative term file.

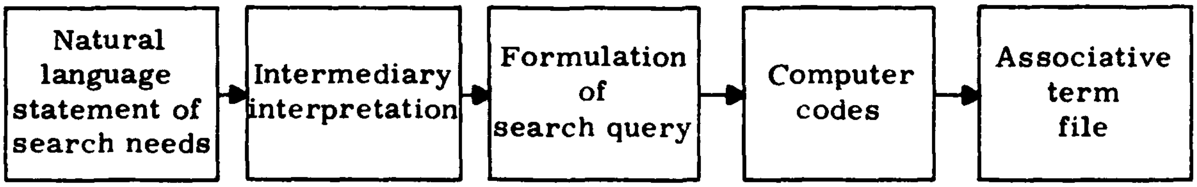
Natural language statement of search needs
Intermediary interpretation
Formulation of search query
Computer codes
Associative term file

Figure 7 Stages of processing in the search subsystem

The processes for each stage may involve a range of equipment and personnel. For example, a user may correspond with an intermediary (if one is used) by letter, by telephone, or in person.

Output subsystems consist of document identification codes translated into document lists (codes, titles, abstracts), intermediary screening of the lists, user screening of the list by document representation and finally, selection of full text in natural language.

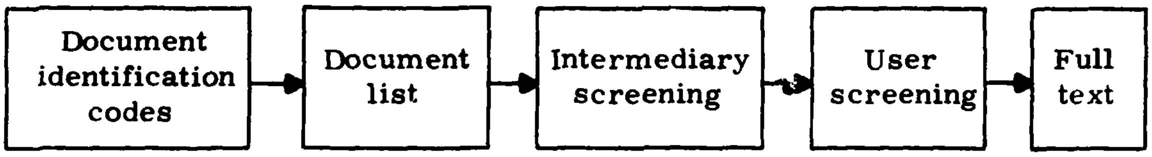

Figure 8 Stages of processing in the output subsystem

Again, an output subsystem may or may not include the stages above. Specific processes may include equipment such as terminal displays and screening could take place on document representations, index terms, or full text with or without intermediaries.

The triad of subsystems is shown in the schema in Figure 9.

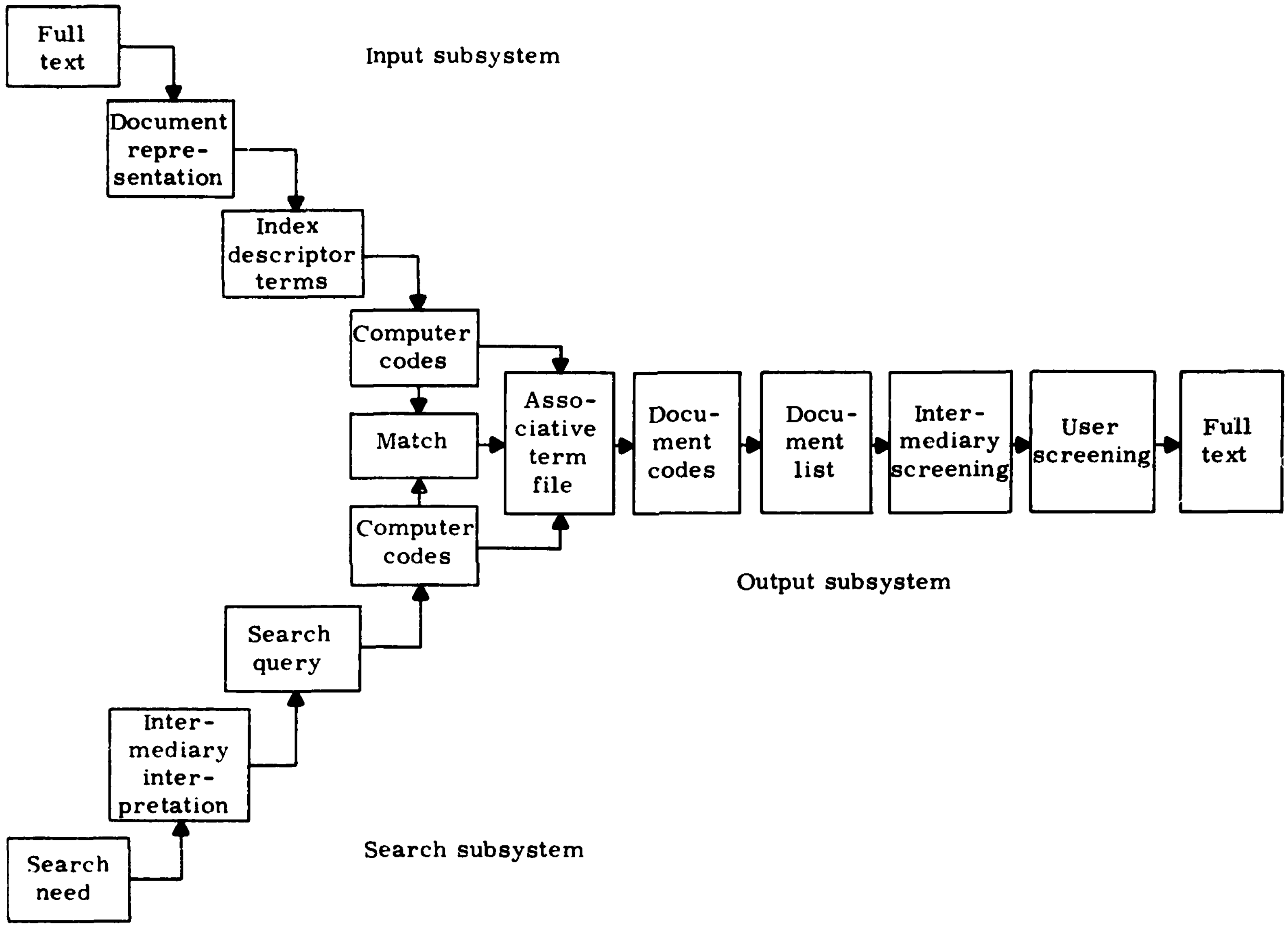

Figure 9 Stages of the triad of input, search, and output subsystems

A basic problem inherent to document retrieval systems concerns matching natural language terminology found in documents with user expressions of search needs. As shown above in Figure 9, traditional input processes distill textual information into computer codes that are matched with similar codes derived through a series of search processes that also distill a natural language statement of search needs into computer codes. Each step is taken in order to reduce costs. However, each step in turn may yield a deterioration in search effectiveness. Therefore, in order to assess a document search and retrieval system, we need to establish a model that determines the cost/effectiveness tradeoff for each step.

Another problem is that the distillation process for input and search may be accomplished independently so that reduction in effectiveness is compounded. A number of techniques are employed to alleviate this problem. For instance, a thesaurus is often used as a common basis for choosing index terms during input and for formulating search queries using terms available in the system. Also, a commonality is achieved in some systems by using the same staff for abstracting and indexing as that used for searching. Thus, the intermediary searcher has the advantage of having a more complete knowledge of the index file and its strengths and weaknesses. Another method of bridging the input/search gap is to have a user provide selected relevant documents that satisfy his needs and use these documents to help formulate queries as well as to determine how effective a search is by observing whether or not the relevant documents are retrieved. A similar technique is to use relevant documents chosen ahead of time or during the first search output to provide a basis for associative term adjustments for subsequent searches. At any rate there are a number of techniques to improve effectiveness by bridging the input/search gap.

The study of retrieval effectiveness described previously involved two input and search systems. Both systems transform the full text into document representations that are then input into computer codes in two fundamentally different ways. These are as follows:

- One process uses a manual indexing system based on humanly assigned index terms selected from a thesaurus. The search file thus created can be searched online or offline using Boolean-type search equations generated from terms chosen from the thesaurus. Searches may be conducted by the user or an intermediary.

- The second process uses a full text input system in which all nontrivial terms are selected from the natural language of a document representation. The search file thus produced can also be searched online or offline using Boolean-type search equations in which terms may or may not be from a controlled list or thesaurus. Again, searches may be conducted by the user or an intermediary.

Search output for both systems in this study were the same. Documents could be identified by computer listing of document number, title, or abstracts depending on search instructions.[7] The first system above involves translating the document natural language into a limited number of descriptors that are available from a thesaurus. Thus, there is a burden on the indexer to choose descriptors from the full text that a user or searcher may envision to satisfy his search needs. The second system, on the other hand, places the burden on the searcher in that he must anticipate all the ways that his search needs can be expressed in the natural language of document text.

Manual indexing using a controlled vocabulary is an expensive process that has administrative problems associated with maintenance of consistency and dependence upon skilled and trained personnel who are particularly difficult to recruit and retain. Under a manual index system the number of concepts identified in a large text is usually somewhat limited so that searches on information in certain portions of a document may never yield the document because it is not indexed under those terms. This is particularly true when the contents of a document are heterogeniousand terms donot lead to other sections of the document. Free-text searching may offer certain economies in input operations. Although, the best economies can be realized when machine-readable information is used for several information by-products such as searching, current awareness, recurring bibliography, and tape sales.

In addition to the subsystems described above and their associated processes, a number of system environmental factors must be considered a part of cost/effectiveness evaluation. Included among such factors are:

- Users - experience, familiarity with the file, number, frequency of use of the system, and search needs (bibliographic, subject, browsing)
- File - number of documents and composition (homogeneity) of subject matter

[7] There is some reason to believe that output screening processes should be chosen based on the combined effectiveness of input, search, and output subsystems. For example, an inexpensive but effective search output screening process could make up for an input system that yields excessive retrieval.

- Documents - size, composition (homogeneity) of subject matter within document, and complexity of language.

Each of these factors contribute to both system cost and effectiveness as will be discussed subsequently.

2. Cost/effectiveness

This section is concerned with cost/effectiveness factors that affect the choice among alternative system designs for document search and retrieval systems generally discussed in the previous section. The complex nature of these systems is highlighted by subsystems that can consist of a number of alternative processes comprised of a myriad of hardware, software, and general procedures. The complexity is enhanced further by the system environment which varies widely from system to system depending on user, file, and document characteristics. In order to cope with these complexities, the systems may be described by a cost/effectiveness model which permits system management to compare many system alternatives without actually developing each alternative.

The model mentioned above incorporates performance effectiveness of the input, search, and output processes and combines all of these measures of effectiveness and their corresponding costs into measures of total system cost, average cost per search, and average cost to retrieve a relevant document. Relevance is defined in terms of the user's judgment as to whether or not a document answers his natural language statement of information need. Every decision concerning the processes above can be made on the basis of the process' contribution to the cost of retrieving relevant documents.

All searches can be performed sequentially to obtain any desired proportion of relevant documents retrieved (recall). Typically, it is found, however, that it becomes increasingly difficult to retrieve each subsequent relevant document. The model can estimate cost/effectiveness at all levels of recall so that management can establish the worth of retrieving all or any portion of relevant documents for a typical search. It is emphasized that there are additional measures of performance effectiveness that must be weighed in decision making such as response time and readactivity. However, the consequences or benefits of decisions can be inferred largely from the value of relevant document retrieval.

The total cost of document search and retrieval systems depends on fixed costs associated with each subsystem (and its processes) and on variable costs related to the number of items input, number of searches conducted, number of items retrieved per search, number of items screened per search, and number of terms in the authority list. The fixed costs are associated with the three subsystems described previously.

- Input - Fixed costs associated with input (C_3) include such items as thesaurus development, staff, keyboard equipment, tape conversion, and update costs.
- Search - Fixed costs associated with searching include computer costs (C_1) such as staff, space rental, computer rental, and fixed computer storage costs. It also includes fixed costs associated with an intermediary or other user/system interface (C_4) such as staff, rent and sundry items.
- Output - Fixed costs associated with screening (C_2) such as rent, staff, storage and display hardware, and other sundry items and fixed costs associated with sending search output to the users (C_5).

Thus, total fixed costs $C' = C_1 + C_2 + C_3 + C_4 + C_5$

The variable costs that are dependent on the file size (X_1) are composed of the cost (C_6) per item of abstracting, indexing, keyboarding, and any other input processing. Also, file loading costs (C_7) include costs that vary with the number of terms (X_5). Thus, the cost component related to file size or input processes is

$$C'' = [C_6 + X_5C_7]\, X_1$$

The variable costs that are dependent on number of searches (X_2) consists of three parts: fixed costs per search, costs dependent on the number of items retrieved (X_3), and costs dependent on the number of items sent to the user (X_4) or those costs generally associated with output. The fixed elements of the search and output costs are the cost of an intermediary per search (C_9) and the set-up costs per search for sending titles to users (C_8). Costs related to number of items retrieved include computer retrieval costs per search per item (C_{10}), printout costs per search per item (C_{11}), and screening costs per search per item (C_{12}). The cost dependent on the number of items sent to users (X_4) is the cost per search per item (C_{13}). The entire search/output component of cost is

$$C''' = [C_8 + C_9 + X_3\,(C_{10} + C_{11} + C_{12}) + X_4C_{13}]\, X_2$$

Thus, the total cost may be expressed as:

$$C = C_1 + C_2 + C_3 + C_4 + C_5 + X_1\,(C_6 + C_7X_5) +$$
$$X_2\,[C_8 + C_9 + X_3\,(C_{10} + C_{11} + C_{12}) + X_4C_{13}]$$

Some systems may not have linear relationships between costs and variables such as number of terms input. However, the general equation can be used in one form or another to compare cost/effectiveness trade-offs of system alternatives if effectiveness measures are available.

As mentioned previously, every process in input, search, and output yields errors. The problem is to characterize these errors and combine them by means of the model so that an overall measure of accuracy can be estimated. The sources of error can be described as a set of conditional probabilities using the following notation:

V_r, relevant with respect to verbalized request;

$V_{\bar{r}}$, nonrelevant with respect to verbalized request;

C_r, relevant with respect to intermediary's interpretation;

$C_{\bar{r}}$, nonrelevant with respect to intermediary's interpretation;

E_r, relevant with respect to encoded request;

$E_{\bar{r}}$, nonrelevant with respect to encoded request;

R_r, relevant with respect to system's response;

$R_{\bar{r}}$, nonrelevant with respect to system's response;

S_r, relevant with respect to screener's judgment; and

$S_{\bar{r}}$, nonrelevant with respect to screener's judgment.

Conditional probabilities are designated by the standard notation $P\ (A/B)$, which is read "the probability of A, given B." Thus, $P\ (C_r/V_{\bar{r}})$ means "the probability that a document is relevant to the intermediary's interpretation, given that it is not relevant to the verbalized request."

Whether one can express relationships among the components of a retrospective searching system as probabilities, and the context within which such probabilities have meaning, requires some elaboration. Let us consider a probability such as $P(R_r/V_r)$, that is, the probability that a document relevant to the verbalized request (V_r) will be retrieved (R_r). If one chooses a request at random from the stream of requests entering the system, presumably it would be possible to say whether any document in the system was relevant or nonrelevant to that request. Also, one can observe whether such a document is retrieved or is not retrieved by the system. The relative frequency with which relevant documents are retrieved by the system should approach stability as the number of observations is increased. Since an observation is identifiable with a document, such stability should occur either if many documents are matched against a single request or if a few documents in each of many searches are matched against their separate search requests. If the ratio generated by the latter method does, in fact, approach stability as the number of requests increases, the value approached as a limit will be referred to as "the probability of retrieval by the system, given relevance to the verbalized request," that is, $P(R_r/V_r)$. In practice, one is always working with relative frequencies, since the limiting values are unknown. It is convenient in model construction, however, to work with the conceptual limits and to call them probabilities.

The relationships can be expressed in a series of effectiveness probabilities as shown below:

	Relevance with respect to intermediary's interpretation	
Relevance with respect to verbalized request	$\bar{C}_r$	C_r
V_r	$P(\bar{C}_r/V_r)$	$P(C_r/V_r)$
$\bar{V}_r$	$P(\bar{C}_r/\bar{V}_r)$	$P(C_r/\bar{V}_r)$
	Relevance with respect to encoded request	
Relevance with respect to intermediary's interpretation	$\bar{E}_r$	E_r
C_r	$P(\bar{E}_r/C_r)$	$P(E_r/C_r)$
$\bar{C}_r$	$P(\bar{E}_r/\bar{C}_r)$	$P(E_r/\bar{C}_r)$
	Relevance with respect to response by system	
Relevance with respect to encoded request	$\bar{R}_r$	R_r
E_r	$P(\bar{R}_r/E_r)$	$P(R_r/E_r)$
$\bar{E}_r$	$P(\bar{R}_r/\bar{E}_r)$	$P(R_r/\bar{E}_r)$
	Relevance with respect to screener's interpretation	
Relevance with respect to verbalized request	$\bar{S}_r$	S_r
V_r	$P(\bar{S}_r/V_r)$	$P(S_r/V_r)$
$\bar{V}_r$	$P(\bar{S}_r/\bar{V}_r)$	$P(S_r/\bar{V}_r)$

Mathematically, a model is constructed that shows:

- the probability that a relevant document will be retrieved (recall)
- the probability that a nonrelevant document will be retrieved (fallout)

From these measures, additional measures can be derived such as precision (proportion of retrieved documents that are relevant) and total retrieval. The model is described as a finite Markov chain with absorbing states.

In the experiments described, input and search errors were observed as a single measure. The sources of error were then diagnosed for a descriptor and full-text system. Effectiveness probabilities for the combined error are given by the following equations:

$$P(R_r/V_r) = P(R_r/C_r)\,P(C_r/V_r) + P(R_r/C_{\bar{r}})\,P(C_{\bar{r}}/V_r)$$

$$P(R_r/V_{\bar{r}}) = P(R_r/C_r)\,P(C_r/V_{\bar{r}}) + P(R_r/C_{\bar{r}})\,P(C_{\bar{r}}/V_{\bar{r}})$$

$$P(S_r, R_r/V_r) = P(S_r/V_r)\,P(R_r/V_r)$$

$$P(S_r, R_r/V_{\bar{r}}) = P(S_r/V_{\bar{r}})\,P(R_r/V_{\bar{r}})$$

The next section gives an example in which the effectiveness probabilities are computed along with corresponding costs to show the cost/effectiveness relationship for a number of system alternatives.

Cost/effectiveness considerations must include all subsystems and their processes. For example, consider abstracts used for full-text searches.

Document Representation (Abstract)

- Function
    - Input subsystem - reduces amount of information input into the system
    - Output subsystem - provides a mechanism for screening search output
    - Other systems - current awareness, reference, recurring bibliographies

- Relationship to effectiveness
    - Input subsystem - increases missed relevant document $P(R_{\bar{r}}/E_r)$, increases retrieved nonrelevant documents $P(R_r/E_{\bar{r}})$.
    - Output subsystem - increases missed relevant documents $P(S_{\bar{r}}/V_r)$, increases retrieved nonrelevant documents $P(S_{\bar{r}}/V_{\bar{r}})$.
- Relationship to cost[8]
    - Input subsystem - fixed costs (C_1, C_3) and variable costs (C_6, C_7).
    - Output subsystem - fixed costs (C_2, C_5) and variable costs (C_8, C_{12}).

Environmental factors also must be considered since they are related to both effectiveness and cost. For example, there is some evidence[9] that the looseness of language such as in the social sciences may result in larger retrieval for full-text searches than in other scientific disciplines. The point is that decisions concerning each process must be considered in terms of its environment and in terms of its affect on cost and effectiveness through the entire system. Some examples of this are given in the next section.

[8] C. P. Bourne, J. B. North, and M. S. Kassan, Abstracting and Indexing Rates and Costs: A Literature Review, ERIC Clearinghouse for Library and Information Sciences, University of Minnesota, Minneapolis, Minnesota, May 1970.

[9] J. Katzer, Large Scale Information Processing Systems: Cost Benefits Analysis, Syracuse University School of Library Science, Syracuse, New York, July 1971.

3. An example

An example is given in this section for cost/effectiveness determination for combinations of two input, three search, and five output processes. Effectiveness results are given from the CIRCOL study and studies conducted by others. Typical costs are given from a study of typical costs observed over a broad range of systems and sources. Examples of system comparisons are given for estimated annual cost, cost per search, cost per document retrieved and cost per relevant document retrieved. These values are also displayed for a range of alternative numbers of searches in order to show the net effect of this variable.

The input processes in this example are descriptor (manual index) input and full-text input of a document representation. The Descriptor (index) System involves humanly assigned index terms selected from a thesaurus and the Full-Text System involves selection of all nontrivial terms from an entire document representation. Both systems invert the file so that computer searches involve identification of a list of document codes from each term.

The search system includes searches by a user and by an intermediary. In the latter case, communication is either written or oral between the user and the intermediary. Searches of the descriptor input file are by a Boolean logic combination of terms chosen from a thesaurus. Searches of the full-text input file are by Boolean logic combination of terms chosen from free language of nontrivial terms. Measures of effectiveness are given for intermediary interpretation. Measures of effectiveness are given at four levels of recall [$P(R_r/C_r)$ at .25, .50, .75, 1.00] for descriptor and full-text searches.

The output system involves the following five combinations of screening processes:

- No screening in which case the user receives the search output directly
- Loose screening by an intermediary on titles and abstracts
- Tight screening by an intermediary on titles and abstracts

- Loose screening by an intermediary on titles and topic tags
- Tight screening by an intermediary on titles and topic tags

Values for estimated costs and effectiveness probabilities are given in Table XVI on the next page.

In the example let:

X_1 = 365,000 total documents in the file
X_2 = 50,000 searches per year
X_3 = number of items retrieved per search
X_4 = number of items sent to users per search
X_5 = 2,200 terms in authority list.

A worked example is given for the model above in Appendix II for a full-text system without an intermediary, searching at 25 percent recall, and with loose screening on titles and abstracts.

There are 120 combinations of three input, five output, and two search processes (at four levels of recall each). Rather than summarize all 120 combinations, a few examples of results are given below for illustrative purposes.

In the example, it was found that no intermediary yielded less cost per relevant document retrieved at all levels of searching and with all screening results. Typical results are given below for tight title and abstract screening at 75 percent recall.

System combination	Annual system cost ($)	Cost/ search ($)	Cost/ document retrieved ($)	Cost/relevant document retrieved ($)
Descriptor Input/Search				
No intermediary	2,092,917	41.86	1.99	1.99
By telephone	3,865,617	77.31	3.68	3.68
By letter	2,146,667	42.93	2.04	2.15
Full-Text Input/Search				
No intermediary	4,334,385	86.69	3.94	4.13
By telephone	4,584,885	91.70	4.16	4.37
By letter	4,397,385	87.95	4.00	4.19

Table XVI Typical costs and effectiveness probabilities for alternative systems

Alternatives	Effectiveness probabilities		Fixed costs		Variable costs		
Intermediary	$P(C_r/V_r)$[a]	$P(C_r/V_{\bar{r}})$[a]	C_4[e]		C_9[e]		
None	0.985	0.0000011	$ 0		$10/search		
By telephone	0.975	0.0000017	500		15/search		
By letter	0.950	0.0000034	500		11.25/search		
Input/search	$P(R_r/C_r)$[b]	$P(R_r/C_{\bar{r}})$[b]	C_1[a]	C_3[a]	$C_{10} + C_{11}$[a]	C_6[a]	C_7[a]
Descriptor	0.25	0.000052	$24,440	$6,925	$0.058/item retrieved	$0.6375/ item input	$0.000038/ item/term
	0.50	0.000126					
	0.75	0.000277					
	1.00	0.000548					
Full-text	0.25	0.000068	$32,940	$7,350	$0.26/item retrieved	$1.675/item input	$0.00019/ item/term
	0.50	0.000164					
	0.75	0.000323					
	1.00	0.000669					
Output (screen)	$P(S_r/V_r)$	$P(S_r/V_{\bar{r}})$	C_5[e]	C_2[e]	C_8[e]	C_{12}[e]	C_{13}[e]
No screen	1.00	1.00	$ 250	$ 0	$0	$0	$0
Loose T&A	0.8127[d]	0.2871[d]		4,000	0.35/item	0.102/item	0.004/item
Tight T&A	0.5405[c]	0.0051[c]		5,000	0.35/item	0.125/item	0.004/item
Loose T&TT	0.3495[d]	0.1858[d]		3,000	0.20/item	0.05/item	0.002/item
Tight T&TT	0.2322[c]	0.0033[c]		4,000	0.20/item	0.08/item	0.002/item

Sources:

[a] APA report

[b] CIRCOL study

[c] P. Atherton unpublished report

[d] Combination of b and c

[e] Judgment

The difference between no intermediary and written requests is so small for the Descriptor System that other factors would probably dictate a decision concerning the use of an intermediary such as availability of users (as well as intermediaries), time, and training. On the other hand, there appears to be little difference in the search system with full-text input/search so that other factors may also be considered.

Also, no screening appears to be the best output system at all levels of recall and with all search systems. Typical results are given for no intermediary and searches at 75 percent recall.

System combination	Annual system cost ($)	Cost/ search ($)	Number documents sent to users	Cost/ document retrieved ($)	Number relevant documents sent to users	Cost/ relevant document retrieved ($)
Descriptor Input/ Search						
Loose T&A	1,934,817	38.70	60	0.65	31	1.25
Tight T&A	2,092,917	41.86	21	1.99	21	1.99
Loose T&TT	1,567,017	31.34	32	0.98	13	2.41
Tight T&TT	1,265,717	25.30	9	2.81	9	2.81
No screen	1,224,817	24.50	140	0.17	39	0.64
Full-Text Input/ Search						
Loose T&A	4,162,585	83.25	65	1.28	31	2.69
Tight T&A	4,334,385	86.69	22	3.94	21	4.13
Loose T&TT	3,751,985	75.04	35	2.14	13	5.77
Tight T&TT	3,979,785	79.60	9	8.84	9	8.84
No screen	3,381,185	67-62	156	0.43	38	1.78

Even though no screening yields less cost per relevant document retrieved, it might be best to choose loose screening on titles and abstracts if the cost of obtaining and screening the false drops by the users is greater than the difference of about $14 and $16 per search observed in the two systems. Also, the 72 and 84 false drops sent by the two systems may discourage users from employing the system.

If one is concerned about the cost of increased quality of searches, he can compare a single system at various levels of recall. An example is given in the following table for no intermediary and no screening.

System combination	Annual system cost ($)	Cost/ search ($)	Total documents retrieved	Cost/ document retrieved ($)	Relevant document retrieved	Cost/ relevant document retrieved ($)
Descriptor Input/ Search						
25 percent recall	900,817	18.02	32	0.56	13	1.39
50 percent recall	1,020,817	20.42	72	0.28	26	0.79
75 percent recall	1,224,817	24.50	140	0.17	39	0.64
100 percent recall	1,557,917	31.16	251	0.12	51	0.61
Full-Text Input/ Search						
25 percent recall	1,823,585	36.47	38	0.96	13	2.81
50 percent recall	2,457,185	49.14	86	0.57	26	1.89
75 percent recall	3,381,185	67.62	156	0.43	38	1.78
100 percent recall	5,215,985	104.32	295	0.35	51	2.05

It would appear that the Descriptor Input/Search System becomes more efficient at increased levels of recall from the cost per relevant retrieved standpoint. The Full-Text Input/Search System goes down and then back up at 100 percent recall. It is important, however, to note that if one considers user costs necessary to screen, the cost relationships may be substantially different. For example, if it costs a user, say, $0.25 per document to go through the search output, the total cost per search and cost per relevant retrieved (assuming the user does not screen out relevant documents) would be as shown in the following table.

System combination	Cost/search ($)	Cost/relevant document retrieved ($)
escriptor Input/ earch		
25 percent recall	26.02	2.00
50 percent recall	38.42	1.48
75 percent recall	59.50	1.53
100 percent recall	93.91	1.84
'ull-Text Input/ earch		
25 percent recall	45.97	3.54
50 percent recall	70.64	2.72
75 percent recall	106.62	2.81
100 percent recall	178.07	3.49

The most efficient search levels in both systems appear to be in the 50 to '5 percent recall ranges.

Another important consideration is the number of searches performed er year. Costs are given on the next page for 10,000; 25,000; 50,000 and 75,000 searches.

The costs over 50,000 searches do not decrease by much. It is clear, owever, in this example that the system loading probably should be over 25,000 in order to make the systems economically feasible.

In all instances of this hypothetical example, Full-Text Input/Search is more expensive than the Descriptor Input/Search. The principal contribution to the cost difference is in the input cost. It is emphasized that these costs might be amortized or allocated partially to other systems. For instance, if the full-text input tapes were used for photocomposition, current awareness or tape sales, the input costs could largely be allocated to these systems and the two system costs might well be more in line with one another.

Finally, one should not blindly accept the costs given in the hypothetical examples given in the next table. Although these costs are considered to be "reasonable" after having conducted a review of costs under another contract. Every system environment will not only require different costs but will also yield somewhat different performance effectiveness as indicated in the previous section.

Table XVII Alternative system costs

Alternative combination	Annual system cost ($)	Cost/ search ($)	Cost/ document retrieved ($)	Cost/ relevant document retrieved ($)
Descriptor System 25 percent recall				
10,000	416,017	41.60	1.30	3.20
25,000	597,817	23.91	0.75	1.84
50,000	900,817	18.02	0.56	1.39
75,000	1,203,817	16.05	0.50	1.23
Full-Text System 25 percent recall				
10,000	1,008,285	100.83	2.65	7.76
25,000	1,313,985	52.56	1.38	4.04
50,000	1,823,585	36.47	0.96	2.81
75,000	2,332,985	31.11	0.82	2.39
Descriptor System 100 percent recall				
10,000	547,417	54.74	0.22	1.07
25,000	926,317	37.05	0.15	0.73
50,000	1,557,917	31.16	0.12	0.61
75,000	2,263,567	30.18	0.12	0.59
Full-Text System 100 percent recall				
10,000	1,686,785	168.68	0.57	3.31
25,000	3,010,235	120.41	0.41	2.36
50,000	5,215,985	104.32	0.35	2.05
75,000	7,421,735	98.96	0.34	1.94

APPENDIX I

EXPERIMENTAL DESIGN AND METHODOLOGY

1. Phase 1

a. Data Collection

For the 30 searches which were chosen by CIRCOL personnel, each intelligence (technical) analyst was asked to fill out one Form A form, statement of need, which was then returned to the search analyst. At the same time the intelligence analyst was asked to fill out one Form B, list of relevant documents, which was to be given either to an administrator or the search analyst provided that the search analyst had previously formulated the search query. Upon receipt of the statement of need (Form A), the search analyst then proceeded to formulate and run the query in the normal manner. Upon completion of the run, the search analyst judged each retrieved document for relevance within these categories: relevant, questionable, not applicable. Copies of these same documents were then given to the intelligence analyst who also judged them within these categories: relevant, useful (and also relevant), not relevant. Once this stage was reached, CIRCOL personnel attempted to obtain copies of as many of the documents listed in Form B (list of relevant documents) as possible. In addition to these copies such information as descriptors, titles, etc. was needed. Problems incurred regarding Form B are discussed in Appendix III. At this point the following material regarding each of the 30 test searches was forwarded to Westat:

- Form A (statement of need) completed by the intelligence analyst.
- Form B (list of relevant documents) completed by the intelligence analyst.
- Copies of Form B documents along with the necessary bibliographic information (i.e., titles, descriptors).
- A copy of the search query and printout obtained by the search analyst.
- Copies of all retrieved documents judged for relevance by the search analyst.
- Copies of all retrieved documents judged for relevance by the intelligence analyst.

b. Procedure for Simulation

(1) Determining Recall

(a) Descriptor System - Indexed input with controlled vocabulary searching

The search query which was formulated and run by the CIRCOL search analyst for a particular search was matched against the corresponding Form B document (list of relevant documents) bibliographic information (descriptors and title) to determine the proportion of those documents which would be retrieved by that particular search. For example, suppose ten documents had been listed by the intelligence analyst on Form B as being relevant to his information need. The search analyst chose the term "aircraft" and qualified by DATE GE 66, i.e., only those documents bearing date of 1966 or later would qualify. Upon examining the descriptors and titles of the ten relevant documents, it is found that six of those contain the word "aircraft" and bear a date of at least 1966. An estimate of the recall ratio (number of relevant documents retrieved/total number of relevant documents) for that particular search is 60 percent. This procedure is followed for all thirty searches yielding a set of recall estimates.

(b) Full-Text Controlled Vocabulary System - Full text input with controlled vocabulary retrieval

Basically the same procedure is used to determine the recall for this system as that used above. The CIRCOL search analyst's search query is used as before; however, instead of being matched against the descriptors and title of the Form B relevant documents, it is matched against the corresponding abstract and title. Recall estimates were determined and average recall was figured in the same manner as that above.

(c) Full-Text Natural Language System - Full-text input with natural language retrieval

This system was divided into three separate search methods for the purpose of evaluation. Each part is treated as a separate system. The basis for this division is the method of natural language query formulation. In the first case (a), Need - the intelligence analyst's statement of need is used to formulate the query. The second case (b), Query - uses the CIRCOL search analyst's query to formulate a natural language query. The third (c), Ideal - uses the documents specified by the intelligence analyst in Form B to formulate the best natural language query. "Best", in this case, refers to the highest feasible recall which results in a reasonable number of documents retrieved. For each of these three search methods the recall values and average recall for the system, (or each subsystem in this case), was determined as for the Full-Text Controlled Vocabulary System by matching each query against the abstract and title of the Form B documents. Average recall was estimated by dividing the sum of relevant retrieved (from Form B) by the sum of the total number of relevant observed (from Form B) where the numerator and denominator are summed over all searches. This estimation equation was used throughout.

(2) Determining Total Number of Documents Retrieved

(a) Descriptor System - Indexed input with controlled vocabulary searching

The procedure used for calculating total number of documents retrieved under this system was merely to count the number retrieved (Z) by each of the 30 CIRCOL search analyst's queries over the 365,000 document database. A simple average was then calculated from the 30 searches.

(b) Full-Text Controlled Vocabulary and Natural Language Systems - Full-text input with controlled vocabulary or natural language retrieval

A different procedure for determining total number of documents retrieved was necessary for both Full-Text Systems since there was not a sufficiently large database containing full-text input. Full-text input began sporatically after the first 365,000 documents were input to the system and full-text input was begun on substantially all documents after 414,000 documents were input to the system. It was not known for sure which documents between 365,000 and 414,000 were input both by the descriptor input, as well as full-text input so that the only known common database was those documents input after 414,000.

The first problem was that there was only a small database common to both Descriptor and Full-Text input Systems. The second problem was that it was not possible to distinguish from search printout whether a document was retrieved by a descriptor, title, or full-text. Thus, it was necessary to look at the entire document of every document retrieved in order to determine whether it was retrieved by descriptor, full-text, or both. Since this process was extremely time consuming it was decided to look at the entire document retrieval for a sample of six searches.

Next we wished to extrapolate estimates of total retrieval for Full-Text searches to the larger database of 365,000 documents so that a direct comparison could be made with descriptor retrieval on this database. Each full-text search on the common database (414,000 to 485,000) yielded documents retrieved by full-text input (x) and by descriptor input (y). The equivalent total retrieval (Y) of descriptor input over the 365,000 and under database was also observed. The question, then was how to estimate the total retrieval of full-text input (X) over the 365,000 and under database. It was assumed that:

$$\frac{X}{x} = \frac{Y}{y}$$

so that

$$X = Y \frac{x}{y}$$

Stated in words, we assumed that the ratio of retrieval from descriptor input from 365,000 and under to retrieval of descriptor input from the common database (414,000 t0 485,000) would be the same as the ratio of retrieval from full-text input from 365,000 and under to retrieval of full-text input from the common database (414,000 to 485,000) if, in fact, there had been a full-text input for the 365,000 and under database. The resultant equation:

$$X = Y \frac{x}{y}$$

is a statistical estimation procedure commonly referred to as "ratio estimation". Note that all of the observed total retrieval (x, y and Y) are from the same full-text search query.

The search prccedure to accomplish the estimates of total retrieval (X) over 30 searches was as follows. A sample six of the 30 searches was chosen randomly[10]. These six searches were run on CIRCOL independently using the full-text search queries. The searches were conducted on the common database (414,000-485,000). Each document retrieved from this database was examined to determine how many were retrieved by full-text input (x) and by descriptor input (y). Each query was then posed to the 365,000 and under database to determine how many documents were retrieved by descriptor input (Y). Total retrieval from full-text input (X) was then estimated by the equation above.

For example, assume that a full-text search query yielded eight documents from the common database (414,000 to 485,000). Examination of the eight entire documents show that six would be retrieved by full-text input (x) and four by descriptor input (y). Assume further that the same search query yielded 20 documents from descriptor input (Y) from the 365,000 and under database. The estimate of full-text input from the latter database is:

$$X = 20 \frac{6}{4} = 30$$

[10] All searches were put in alphabetical order and a 20 percent sample (six searches) was chosen by random intervals of five starting the third search. (Search numbers chosen were 3, 8, 13, 18, 23, 28).

The same procedure was used for the remaining 24 test searches except that only Y was observed and (x/y) was found from the average ($\Sigma x/\Sigma y$) over the six searches described above. Estimates for the standard error of this estimate are given in the next section.

In summary, first, ratio estimates were found for each of the 24 sample searches. The average ratio of the six sample searches under each system was determined. This average ratio was then multiplied by the number of documents retrieved by the corresponding search under the present system - 30 times for each system. This yielded an estimate of the retrieval size under each system for the 30 searches involved. These may then be averaged as was done for the first system.

In addition to determining recall and total retrieval values for the three systems, the original search strategies were broadened and narrowed to gain a better understanding and feel for the CIRCOL system. Each term in the particular strategy was taken separately and all related terms, or a sample of all related terms in cases of exceptionally large numbers was recorded. Various combinations, in the same basic format as the original query wherever possible, were recorded and run at the Library of Medicine (NLM) terminal in order to compare the number of documents retrieved by broader or narrower terms.

2. Phase 2

a. System Comparison over Four Levels of Recall

This portion of Phase 2 consisted of a comparison of two systems (Descriptor and Full-Text Natural Language) in terms of recall (four levels) and total number of documents retrieved, expressed both as an absolute figure and as the number of documents that must be scanned in order to obtain a particular level of recall. As an added consideration, precision and fallout measures were also determined.

Treating the two systems separately, four search queries were formulated for each of 24 of the 30 Phase 1 sample searches. Six of the sample searches were discarded due to lack of a sufficient number of known relevant documents or to the technical nature of the request. Each of the queries was designed to attain one of the four required levels of recall (25 percent, 50 percent, 75 percent, 100 percent).

The determination of the number of documents retrieved by each of these 192 searches (24 topics x 4 levels of recall x 2 systems) was different for each of the two systems. Under both systems, however, retrieval was limited to the first 365,000 documents of the file. This was done to obtain a retrieval number limited definitely to descriptor retrieval since the remainder of the file (approximately 135,000 documents) contains a mixture of descriptors (topic tags), abstracts, and titles, all of which are searched without the ability to discriminate on control fields.

(1) Descriptor System

For the Descriptor System searches the query was run on the first 365,000 documents of the file[11] and the number retrieved was noted.

(2) Natural Language Full-Text System

For the full-text searches the query was run on the first 365,000 document database as for the descriptor searches. This number (Y) represented the retrieval size based on a natural language search of the descriptor (topic tag) field and had to be converted to a comparable figure based on the full-text (abstract) field. This was done by using a ratio estimating procedure as in Phase 1. The common document database was used which contained only documents for which both descriptors (topic tags) and abstracts are contained in the system and were therefore both searchable.

[11] This is accomplished by searching on the entire file then searching the portion of the file that is beyond (later than) 365,000. The latter number is then subtracted from the former number, thus yielding the number of documents retrieved from the first 365,000 documents.

A sample of five of the Full-Text System search queries over the four levels of recall, 20 queries, was run on this database and printouts of the results were ordered. Each printout was read separately and the number of documents retrieved for each search by descriptors (topic tags) (y) was recorded. The same procedure was then used on the same printouts and the number of documents retrieved for each search by free-text searching (abstracts) (x) was recorded. These two figures provided the ratio of abstract (x) to descriptor (y) retrievals or x/y. When combined with the ratio obtained in Phase 1 and when multiplied by Y (descriptor retrieval over the larger database of 365,000) this ratio yielded an estimate of full-text natural language retrieval size that is based on easily obtainable descriptor retrieval size.

The same procedure was then used to modify recall and retrieval figures for the Phase 1 descriptor and full-text queries in order to make them consistent with the corresponding Phase 2 queries. Modification of the Phase 1 queries was necessary because the two sets of queries were originally run on a different size recall base. Following completion of this procedure, recall and retrieval figures were then available for five queries under each of the two systems (Descriptor and Full-Text) for each search. An example is given in Appendix II.

Next, five queries (for descriptor and full-text searches) were arranged in their likely order of selection as though they were searched sequentially. Thus, a series of queries were available upon which to determine total retrieval at the four levels of recall (.25, .50, .75, and 1.00). It was decided to interpolate where in a list of retrieved documents a relevant document would be retrieved. Interpolation is made by the following equation[12, 13]:

$$\frac{K(n+1)}{k+1} + N, \text{ where:}$$

k = total number of relevant documents estimated to be retrieved in the set of retrieved documents

n = total retrieved set of documents

K = the order of relevant documents (K = 1, 2, . . . , k)

N = cumulative retrieval prior to the current set

[12] "Potential Improvement of Retrieval by Associative Adjustment of the File." Bryant and King, p. 6 Westat Research, Inc.

[13] Example in Appendix II.

Assume that the first query yielded 10 documents retrieved (N) with none relevant from the list of Form B documents and the second query yielded 20 documents retrieved (n) and two relevant documents (k) found from Form B. If one were scanning through the printout, the relevant documents would be estimated to be found in the set as follows:

1st relevant document -

$$\frac{K(n+1)}{k+1} + N = \frac{1(21)}{3} + 10$$

$$= 17$$

2nd relevant document -

$$\frac{K(n+1)}{k+1} + N = \frac{2(21)}{3} + 10$$

$$= 24$$

This estimation procedure was followed to estimate when each of the four relevant documents would be discovered which yields an estimate of total retrieval at all four recall levels.

The total number of relevant documents in the file for each of the 24 requests had been calculated under Phase 1 analysis. The number of relevant documents retrieved was calculated by multiplying this number by the recall. By subtracting the number of relevant documents retrieved from the total number of documents retrived, the number of nonrelevant documents retrieved was found. Thus, fallout and precision were determined in the usual manner.

This procedure allowed the comparison of the two basic systems in question over the same levels of recall. The two systems were then directly compared in the same terms. Phase 1 analysis led to the formulation of several hypotheses concerning the two systems. This portion of Phase 2 enabled the further testing of those hypotheses and the formulation of additional ones that were based on more operational data. The effects of broadening and narrowing the search query (simulated through the various levels of recall) were analyzed and the results were compared for the two systems.

b. Recall Failure Analysis

This portion concerned failure analysis to determine the primary sources of failure (nonretrieval of relevant documents) and to make recommendations for improvement within the two systems in question.

The known relevant documents for the previous 30 searches were examined against the search queries under the two systems in question in order to categorize recall failures into one of the following groups:

Descriptor System

- Indexing language
- Indexing process
- Searching process
- User/system interaction

Other

Full-Text System

- Synonym failures
- Searching process
- User/system interaction
- Other

Analysis enabled the primary sources of recall failure to be established for each of the two systems in question.

The difference in number of occurrences between the two systems (Descriptor and Full-Text Natural Language) is caused by a difference in the number of relevant documents which were missed for each system. There were three cases (two Descriptor failures and one Full-Text failure) that could not be assigned to any category and therefore were excluded.

c. Effects of Analyst Experience

The two factors to be examined were: analysis of the effect of technical analyst experience with the system and the use of known relevant documents on system performance.

Four technical analysts (two experienced and two inexperienced) were given five topics (chosen and approved by the group in advance) and asked to identify between them at least six relevant documents. Each analyst was then asked to formulate a natural language query for each of the five topics (using a terminal if desired). Following completion of these queries each analyst was asked to take copies of two of the prespecified relevant documents and see if he could improve his previous query or formulate a better new one using these two relevant documents as a guide. The same two relevant documents were used by each analyst. These sets of queries (two for each analyst for each of the five topics) were then run for total documents retrieved. The same ratio estimating procedure as described under Part A was used to estimate Full-Text retrieval. Recall was calculated using the four remaining prespecified relevant documents after excluding those used during searching.

This part enabled comparison of the performance of both inexperienced and experienced analysts and gave insight into possible improvements in performance by the utilization of relevant documents in search strategy formulation.

d. Analysis of the Effect of File Additions

This part involved analysis of some of the possible effects of the addition of such files as Word Form Conversion and Synonym/Equivalent.

Queries for the two systems were developed under Phase 1. Thirty searches were expanded to simulate the effects of various file additions or modifications. A sample of relevant documents from the original was matched against these new expanded queries and new recall estimates were determined. These queries were then run to determine the total number of documents retrieved using the same ratio estimating procedure described in Part A. All queries (including the original set of queries from Phase 1) were run on a 365,000 document database.

This part gave much needed insight into the effects of adding or modifying various files in terms of changes to both recall and number of documents retrieved.

The Word Form Conversion (WFC) File simulation included precise synonyms, variant spellings, common misspellings, common abbreviations and a few specified endings for both verbs and nouns (i.e., absorber, absorbing). The Synonym/Equivalent File simulation included acronyms, multiple word synonyms, near synonyms, narrower terms and related terms. These items were gathered through use of the CIRC Thesaurus, Webster's Dictionary, Roget's Thesaurus, CIRCOL Substitution List, and portions of another FTD contractor's report concerning file changes (still in progress). An example is given in Appendix IV.

3. Statistics

Total retrieval for Descriptor System searches was estimated as follows:

$\bar{Z} = \frac{\Sigma z}{n}$ where z is total retrieval for descriptor searches from 365,000 and under database
n is the number of searches

Standard error: $\mathrm{Var}(\bar{Z}) = \sqrt{\frac{\Sigma(z_i - \bar{z})^2}{n(n-1)}}$

Total retrieval for Full-Text System searches was estimated as follows:

$$\bar{X} = \frac{\Sigma Y \frac{\Sigma x}{\Sigma y}}{n}$$

where Y is total retrieval for full-text searches from 365,000 and under database

x is total retrieved for full-text searches from full-text input from common database (414,000-485,000)

y is total retrieved for full-text searches from descriptor input from common database (414,000-485,000)

Y is summed over n searches

x and y are summed over 26 queries (5 searches over 4 levels of recall plus 6 from phase 1)

$x/y = p$

Standard error from:

$$\text{Relvar}(\bar{X}) = \text{Var}(\bar{Y}) + \text{Var}(p)$$

$$= \text{Var}(\bar{Y}) + \text{Var}(\bar{x}) + \text{Var}(\bar{y}) - 2\,\text{Covar}(\bar{x}, \bar{y})$$

$$= \frac{\Sigma(Y_i - \bar{Y})^2}{n(n-1)\,\bar{Y}^2} + \frac{\Sigma(y_i - px_i)^2}{m(m-1)\bar{y}^2}$$

$$\text{Var}(\bar{X}) = \bar{Y}^2 (p)^2 \; \text{Relvar}(\bar{X})$$

All variances for $\bar{Y}$ in Phase 1 and Phase 2, Parts A and E, were calculated by the technique outlined on page 108 of Experimental Statistics by Cochran and Cox, John Wiley & Sons, Inc., 1957, Canada. Searches are considered replications and four recall levels are treatments in a design commonly referred to as a randomized block design. The variance Var ($\bar{Y}$) was calculated by the Error Mean Square from analysis of variance at each level of recall.

All Standard Error estimates for Phase 1 and Phase 2, Parts A & E, recall were calculated by:

$$SE = \sqrt{\frac{\Sigma y^2 - 2(p)\Sigma xy + (p)^2 \Sigma x^2}{n(n - 1)x^2}}$$

where y = number of relevant items retrieved

x = total number of relevant items

$p = \frac{\Sigma y}{\Sigma x}$ (recall)

Results of Standard Error calculations are shown below:

	Recall	Retrieval
Phase 1		
System 1	.081	9.4
System 2	.045	----
System 3	.083	----
System 4	.063	21.6
System 5	.059	----
System 6	.044	----
Phase 2		
Part A		
Descriptor System		
25 percent recall		12.0
50 percent recall		8.5
75 percent recall		17.3
100 percent recall		18.3
Full-Text System		
25 percent recall		11.8
50 percent recall		5.6
75 percent recall		14.7
100 percent recall		22.4
Part E		
Descriptor System		
Original	.084	8.6
WFC	.078	8.7
WFC and SE	.084	12.5
Full-Text System		
Original	.105	24.5
WFC	.063	30.6
WFC and SE	.071	70.8

Phase 2, Part D. The analysis of variance described below can be employed to test for the existence of effects due to the different experimental factors.

Table of square roots of retrieval counts

Analyst (i)	Query (k)	Topic (j)				
		1	2	3	4	5
1	1	7.3485	0.0	0.0	1.7321	1.0
	2	0.0	0.0	1.7321	1.4142	12.0416
2	1	2.6458	2.2361	8.5440	11.3578	2.0
	2	2.6458	2.2361	8.5440	11.3578	2.0
3	1	3.3166	2.8285	7.4162	5.2915	6.0828
	2	2.8285	3.7417	4.2426	5.2915	6.0828
4	1	13.7840	8.2462	7.6158	9.7468	16.6433
	2	13.7840	8.2462	7.6158	11.4018	16.6433

Query totals; 20 totals of 2

$$Y_{..k} = \sum_i \sum_j y_{ijk}$$

Analyst totals; 4 totals of 10

$$Y_{i..} = \sum_j \sum_k y_{ijk}$$

Topic totals; 5 totals of 8

$$Y_{.j.} = \sum_i \sum_k y_{ijk}$$

(y_{ijk} = observation in table above on the kth query for the ith analyst on the jth topic)

Grand total; 1 total of 40

$$Y_{...} = \sum_i \sum_j \sum_k y_{ijk}$$

First query x second query totals; 2 totals of 20

$$Y_{ij.} = \sum_k y_{ijk}$$

Query x analyst totals; 8 totals of 5

$$Y_{.k} = \sum_j y_{ijk}$$

Query x topic totals; 10 totals of 4

$$Y_{.jk} = \sum_i y_{ijk}$$

Uncorrected sums of squares

Total uncorrected sum of squares

$$S_T = (1/1)(y_{111}^2 + y_{121}^2 + \ldots\ldots + y_{452}^2)$$

Total uncorrected topic by analyst cells sum of squares

$$S_{TpA} = (1/2)(Y_{11.}^2 + Y_{12.}^2 + \ldots\ldots + Y_{45.}^2)$$

Total uncorrected analyst sum of squares

$$S_A = (1/10)(Y_{1..}^2 + Y_{2..}^2 + Y_{3..}^2 + Y_{4..}^2)$$

Total uncorrected topic sum of squares

$$S_{Tp} = (1/8)\ (Y_{.1.}^2 + Y_{.2.}^2 + Y_{.3.}^2 + Y_{.4.}^2 + Y_{.5.}^2)$$

Total uncorrected query sum of squares

$$S_Q = (1/20)(Y_{..1}^2 + Y_{..2}^2)$$

Total uncorrected analyst by query cells sum of squares

$$S_{AQ} = (1/5) \sum_{i}\sum_{k} Y^2_{i.k}$$

Total uncorrected topic by query cells sum of squares

$$S_{TpQ} = (1/4) \sum_{i}\sum_{k} Y^2_{.jk}$$

Correction factor

$$S_c = (1/40)\,(Y^2_{...})$$

<u>Corrected sums of squares</u>

Total corrected sum of squares

$$SS_T = S_T - S_c \qquad 39 \text{ d.f.}$$

Total corrected topic sum of squares

$$SS_{Tp} = S_{Tp} - S_c \qquad 4 \text{ d.f.}$$

Total corrected analyst sum of squares

$$SS_A = S_A - S_c \qquad 3 \text{ d.f.}$$

Total corrected query sum of squares

$$SS_Q = S_Q - S_c \qquad 1 \text{ d.f.}$$

Interaction of topics and analysts

$$SS_{TpA} = S_{TpA} - C - SS_{Tp} - SS_A \qquad 12 \text{ d.f.}$$

Interaction of topics and queries

$$SS_{TpQ} = S_{TQ} - C - SS_{Tp} - SS_Q \qquad 4 \text{ d.f.}$$

Interaction of analysts and queries

$$SS_{AQ} = S_{AQ} - C - SS_{A} - SS_{Q} \qquad 3 \text{ d.f.}$$

Interaction of topic and analysts and queries

$$SS_{ATpQ} = SS_{T} - SS_{A} - SS_{Tp} - SS_{Q} - SS_{ATp}$$

$$-SS_{AQ} - SS_{TpQ} \qquad 12 \text{ d.f.}$$

Corrected sums of squares for experience

$$SS_{E} = (1/20)((Y_{1..} + Y_{2..})^{2} + (Y_{3..} + Y_{4..})^{2} - C)$$

Corrected sums of squares within-experience

$$SS_{EW} = SS_{A} - SS_{E} \qquad 1 \text{ d.f.}$$

Corrected sums of squares for query by experience level

$$SS_{QE} = (1/10)(T^{2}_{Q1,E1} + T^{2}_{Q2,E1} + T^{2}_{Q1,E2} + T^{2}_{Q2,E2}$$

$$- C - SS_{Q} - SS_{E}) \qquad 1 \text{ d.f.}$$

Corrected sums of squares for query by within-experience level

$$SS_{QEW} = SS_{QA} - SS_{QE} \qquad 2 \text{ d.f.}$$

$T_{Q1,E1} = Y_{1.1} + Y_{2.1}$ (totals of 10)

$T_{Q2,E1} = Y_{1.2} + Y_{2.2}$ (totals of 10)

$T_{Q1,E2} = Y_{3.1} + Y_{4.1}$ (totals of 10)

$T_{Q2,E2} = Y_{3.2} + Y_{4.2}$ (totals of 10)

Analysis of Variance Table

Source	Degrees of Freedom	Sum of Squares	Mean Squares
Whole Plot Analysis			
Topic	4	SS_{Tp}	M_{Tp}
Analysts	3	SS_{A}	M_{A}
Topic by analyst (Error a)	12	SS_{TpA}	M_{TpA}
Subplot Analysis			
Query	1	SS_{Q}	M_{Q}
Query by analyst	3	SS_{QA}	M_{QA}
Query by topic*	4	SS_{QTp}	M_{QTp}
Analyst by topic by query* (Error b)	12	SS_{ATpQ}	M_{ATpQ}

F_o statistic for testing analyst

$$F_o = M_A / M_{TpA}$$

F_o statistic for testing query

$$F_o = M_Q / \text{Error b}$$

F_o statistic for testing query by analyst

$$F_o = M_{QA} / \text{Error b}$$

$M_{Tp} = SS_{Tp}/4$

$M_A = SS_A/3$

etc.

* These probably should be combined.

APPENDIX II

SAMPLE PROCEDURES AND QUERIES

1. Sample Queries, Phase 1

```
#1 Descriptor System
     L1 COMPUT($), COORDINAT($)
     L2 CENTER, FACILITY
     L3 L1 & L2(+1) & SPACE
     IF CYNTUSSR EQ                  Retrieval                 38

#2 Ideal Descriptor System
     L1 COMPUT($)
     L2 L1 & CENTER
     L3 RESEARCH & FACILITY
     L4 L2, L3
     L5 L4 & SPACE
     IF CNTYUSSR EQ Y                Retrieval                472

#3 Full-Text Controlled Vocabulary System
     Same query as for System #1     Retrieval       38 x .52 = 20

#4 Full-Text Natural Language System Based on Need
     L1 COMPUT($), COORDINAT($)
     L2 L1 & CENTER(+1) & SPACE
     IF CNTYUSSR EQ Y                Retrieval      40 x 1.19 = 48

#5 Full-Text Natural Language System Based on Query
     COMPUT($) & CENTER
     IF CNTYUSSR EQ Y                Retrieval       54 x .87 = 47

#6 Full-Text Natural Language System Based on Ideal
     L1 COMPUT($)
     L2 L1 & CENTER(+1) & SPACE
     L3-UR-, SOVIET[14]
     L4 L2 & L3                      Retrieval       60 x .95 = 58
```

[14] Because CIRCOL cannot search on -UR-, this query was actually run three times; once containing the term SOVIET, once qualified by CNTYUSSR EQ Y and once with both. Retrieval shown is the sum of the first and second query retrievals, less retrieval on the third query.

2. Sample Queries, Phase 2, Part A

Analyst's statement of information need:

"Translated documents concerning both radar and radar theory are very valuable in our work. The various types of books include but should not be limited to antenna theory. 1967 or more recent."

a. Descriptor System Queries

#1 Original Query (from Phase 1)

Retrieves relevant documents . . . . BCD

RADAR		
IF DATATYPE EQ T		
AND DATE GE 66	Retrieval	34
AND DPSNR GE 0000365000	Retrieval	5
Retrieval on 365,000 document database		29

Queries Formulated to Simulate Four Levels of Recall

#2 25 Percent Recall

Retrieves relevant document C

RADAR & THEORY(+1)	Retrieval	5
IF DPSNR GE 0000365000	Retrieval	2
Retrieval on 365,000 document database		3

#3 50 Percent Recall

Retrieves relevant documents . . . . CD

RADAR & SIGNAL(+1) & TRACKING		
IF DATATYPE EQ T	Retrieval	3
AND DPSNR GE 0000365000	Retrieval	0
Retrieval on 365,000 document database		3

#4 75 Percent Recall

Retrieves relevant documents . . . . BCD

RADAR & SIGNAL(+1)		
IF DATATYPE EQ T	Retrieval	22
AND DPSNR GE 0000365000	Retrieval	2
Retrieval on 365,000 document database		20

#5 100 Percent Recall

Retrieves relevant documents . . . . . . ABCD

RADAR		
IF DATATYPE EQ T	Retrieval	182
AND DPSNR GE 0000365000	Retrieval	8
Retrieval on 365,000 document database		174

b. Full-Text System Queries

#1 Original Query from Phase 1

Retrieves relevant documents . . . . BC

RADAR		
AND SIGNAL, ANTENNA, THEORY		
IF DATATYPE EQ T		
AND DATE GE 67	Retrieval	10
AND DPSNR GE 0000365000	Retrieval	4
Retrieval on 365,000 document database		6 x 1.54 = 9

Queries Formulated to Simulate Four Levels of Recall

#2 25 Percent Recall

Retrieves relevant documents C

RADAR & SIGNAL & THEORY		
& ANTENNA	Retrieval	5
IF DPSNR GE 0000365000	Retrieval	2
Retrieval on 365,000 document database		3 x 1.54 = 5

#3 50 Percent Recall

Retrieves relevant documents . . . . BC

RADAR & SIGNAL & THEORY		
IF DATATYPE EQ T	Retrieval	6
AND DPSNR GE 0000365000	Retrieval	2
Retrieval on 365,000 document database		4 x 1.54 = 6

#4 75 Percent Recall

Retrieves relevant documents . . . . BCD

RADAR & SIGNAL		
IF DATATYPE EQ T	Retrieval	45
AND DPSNR GE 0000365000	Retrieval	4
Retrieval on 365,000 document database		41 x 1.54 = 63

#5 100 Percent Recall

Retrieves relevant documents . . . . . . ABCD

RADAR		
IF DATATYPE EQ T	Retrieval	182
AND DPSNR GE 0000365000	Retrieval	8
Retrieval on 365,000 document database		174 x 1.54 = 268

Order of preference of running queries based upon subjective decision that query would be most likely to fulfill information requirement stated by analyst.

Descriptor System Queries	#2, #3, #4, #1, #5
Full-Text System Queries	#2, #3, #4, #1, #5

(These two sequences are not always the same for the two systems.)

3. Sample estimation of progressive number of documents scanned at each level of recall - adjusted retrieval for Phase 2, Part A

Query number	Retrieval per query	n	N	Relevant documents retrieved	Number of documents scanned	Recall level achieved %
				Descriptor System		
2	3	3		C	4/2 = 2	25
3	3	3	3	CD	4/2 + 3 = 5	50
4	20	20	6	BCD	21/2 + 6 = 17	75
1	29			BCD		
5	174	203	26	ABCD	204/2 + 26 = 128	100
				Full-Text System		
2	5	5		C	6/2 = 3	25
3	6	6	5	BC	7/2 + 5 = 9	50
4	63	63	11	BCD	64/2 + 11 = 43	75
1	9			BC		
5	268	277	74	ABCD	278/2 + 74 = 213	100

4. Examples Phase 2, Part B

a. Effectiveness equation example alternative combination 111 (Full-Text System):

$$
\begin{aligned}
&P(R_r/V_r) = P(R_r/C_r)P(C_r/V_r) + P(R_r/C_{\bar{r}})P(C_{\bar{r}}/V_r) \\
&P(R_r/V_{\bar{r}}) = P(R_r/C_r)P(C_r/V_{\bar{r}}) + P(R_r/C_{\bar{r}})P(C_{\bar{r}}/V_{\bar{r}}) \\
&P(S_r, R_r/V_r) = P(S_r/V_r)P(R_r/V_r) \\
&P(S_r, R_r/V_{\bar{r}}) = P(S_r/V_{\bar{r}})P(R_r/V_{\bar{r}})
\end{aligned}
\qquad (1)
$$

Sample effectiveness figures used for a Full-Text Natural Language System with no intermediary, searching at 25 percent recall, with loose screening on titles and abstracts.

$P(C_r/V_r) = .985$

$P(C_r/V_{\bar{r}}) = .0000011$

$P(C_{\bar{r}}/V_r) = .015$

$P(C_{\bar{r}}/V_{\bar{r}}) = .9999989$

$P(R_r/C_r) = .25$

$P(R_r/C_{\bar{r}}) = .000068$

$P(S_r/V_r) = .8127$

$P(S_r/V_{\bar{r}}) = .2871$

$P(R_r/V_r) = .246001$

$P(R_r/V_{\bar{r}}) = .000068$

$P(R_r/V_r) = .25 \times .985 + .000068 \times .015 = .246001$

$P(R_r/V_{\bar{r}}) = .25 \times .0000011 + .000068 \times .9999989 = .000068$

$P(S_r, R_r/V_r) = .8127 \times .246001 = .1999$ = recall

$P(S_r R_r/V_{\bar{r}}) = .2871 \times .000068 = .00002$ = fallout

.1999 (recall) x 52 (estimated number of relevant documents in CIRCOL) = 10 = number of relevant retrieved

.00002 (fallout) x 364948 (estimated number of nonrelevant documents in CIRCOL per search) = 7 - number of nonrelevant retrieved

10 + 7 = 17 = total documents retrieved

10 (relevant retrieved)/17 (total retrieved) = .59 or 59% = precision

b. Cost Equation Example

$$C = C_1 + C_2 + C_3 + C_4 + C_5 + X_1 (C_6 + C_7 X_5) + X_2 [C_8 + C_9 + X_3 (C_{10} + C_{11} + C_{12}) + X_4 C_{13}] \qquad (2)$$

Assuming sample cost figures are as follows for a Full-Text Natural Language System with no intermediary, searching at 25 percent recall, with loose screening on titles and abstract:

C_1 =	$32,940	C_{10} & C_{11} =	$ 0.26
C_2 =	4,000	C_{12} =	0.102
C_3 =	7,350	C_{13} =	0.004
C_4 =	0		
C_5 =	250	X_1 =	365,000
C_6 =	1.675	X_2 =	50,000
C_7 =	0.00019	X_3 =	38
C_8 =	0.35	X_4 =	17
C_9 =	10.00	X_5 =	2,200

then:

$$C = 32,940 + 4,000 + 7,350 + 0 + 250 + 365,000 (1.675 + .00019 \times 2,200) + 50,000 [.35 + 10.00 + 38 (.26 + .102) + 17 \times .004]$$

$$= 44,540 + 365,000 \times 2.093 + 50,000 \times 24.18$$

= $2,017.485 = total system cost

$2,017.485/50,000 (number of searches per year) = $40.35 = cost per search

$40.35/17 (number of documents retrieved) = $2.37 = cost per item retrieved

$40.35/10 (number of relevant retrieved) = $4.04 = cost per relevant item retrieved.

5. Sample Phase 2, Part C

Analyst's statement of information need:

"Cooling of various components in aerospace propulsion systems is termed 'thermal management'. At present the majority of work applicable to this field is found in complex theoretical work which generally is only slightly related to propulsion per se. A small amount of material exists, however, which covers the application of some of this theoretical work to particular types of propulsion. It is imperative that this material be found so as to allow the analysts to be able to search out only specific areas and personalities in the theoretical material. Of particular interest is cooling of components in liquid rocket, solid rocket and ramjet engines. Not turbine engines".

Query: The same query was used for both the Descriptor System and the Full-Text System.

L1 PROPULSION
L2 COOL(+1)
L3 TURBINE & ENGINE(+1)
L4 L1 & L2 & L3(NOT)

An examination of the relevant documents not retrieved by this query revealed that the titles and topic tags contained such terms as PROPELLANT, rather than PROPULSION, HEAT TRANSFER or HEAT EXCHANGE rather than COOL($), and COMBUSTION rather than PROPULSION & COOL($).

The Descriptor System query retrieved one out of seven relevant documents. The use of the query term COOL($) alone was considered a Searching failure to cover all reasonable approaches, due to the fact that neither 'heat transfer' nor 'heat exchange' were searched as alternative descriptors. Other failure categories involved under the Descriptor System were Index Language (endings), Indexing (lack of exhaustivity) and Searching (formulation too specific).

The Full-Text System query retrieved none of the seven relevant documents. The omission of PROPELLANT in the query was considered a Synonym failure to consider a related term. Other failure categories involved were Searching (failure to cover all reasonable approaches and formulation too specific) and User/System Interface (request more specific than actual need).

6. Sample Queries, Phase 2, Part D

Topic: "What are foreign developments in the area of hybrid propulsion system?"

Initial Query: (inexperienced analyst)

```
    HYBRID & PROPELLANT                 Retrieval   129
```

Modified Query: none

Initial Query: (experienced analyst)

```
    L1 HYBRID
    L2 PROPULSION, ENGINE, MOTOR
    L3 L1 & L2
    IF SUBJCODE SC '21'                 Retrieval    95
```

Modified Query:

```
    L1 HYBRID, LITHERGOL
    L2 PROPULSION, PROPELLANT, ENGINE, MOTOR, ROCKET
    L3 L1 & L2
    IF SUBJCODE SC '21'                 Retrieval   130
```

7. Sample Queries, Phase 2, Part E

a. Descriptor System

Original query from Phase 1, System #1

```
ALUMINUM & ALLOY(+1) & LITHIUM
IF CYNTUSSR EQ Y                              Retrieval   33
AND DPSNR GE 0000365000                       Retrieval   10
             Retrieval on 365, 000 document database      23
```

Same query expanded to simulate use of Word Form Conversion File

```
L1 ALUMINUM, ALUMINIUM, AL
L2 ALLOY, AMALGAM
L3 LITHIUM, LI
L4 L1 & L2(+1) + L3
IF CYNTYUSSR EQ Y                             Retrieval   34
AND DPSNR GE 0000365000                       Retrieval   11
             Retrieval on 365, 000 document database      23
```

Same query expanded to simulate use of Synonym/Equivalent File

```
L1 ALUMINUM, ALUMINIUM, AL
L2 ALLOY, AMALGAM
L3 L1 & L2(+1)
L4 L1 & BASE(+1) & L2(+1)
L5 L1 & CONTAINING(+1) & L2(+1)
L6 SINTERED & L1(+1) & POWDER(+1)
L7 L1 & SOLDER(+1)
L8 L1 & INTERMETALLIC(+1) & COMPOUND(+1)
L9 L1 & METALLURGY(+1)
L10 L3, L4, L5, L6, L7, L8, L9
L11 LITHIUM, LI
L12 L11 & L10
IF CYNTUSSR EQ Y                              Retrieval   59
AND DPSNR GE 0000365000                       Retrieval   16
             Retrieval on 365, 000 document database      43
```

b. Full-Text System

Original query from Phase 1, System #4

```
ALUMINUM & ALLOY & LITHIUM
OR AL & LI & ALLOY
```

```
IF CNTYUSSR EQ Y                          Retrieval           65
AND DPSNR GE 0000365000                   Retrieval           19
         Retrieval on 365,000 document database   46 x 1.54 = 71
```

Same query expanded to simulate use of Word Form Conversion File

```
L1 ALUMINUM, ALUMINIUM, AL
L2 ALLOY, AMALGAM
L3 LITHIUM, LI
L4 L1 & L2 & L3
IF CYNTUSSR EQ Y                          Retrieval           66
AND DPSNR GE 0000365000                   Retrieval           20
         Retrieval on 365,000 document database   46 x 1.54 = 71
```

Same query expanded to simulate use of Synonym/Equivalent File

```
L1 ALUMINUM,
L2 ALLOY, AMALGAM, SOLDER, METALLURGY
L3 LITHIUM, LI
L4 L1 & L2
L5 SINTERED & L1(+1) & POWDER(+1)
L6 L1 & INTERMETALLIC & COMPOUND(+1)
L7 L4, L5, L6
L8 L3 & L7
IF CYNTYUSSR EQ Y                         Retrieval           69
AND DPSNR GE 0000365000                   Retrieval           21
          Retrieval on 365,000 document database  48 x 1.54 = 74
```

APPENDIX III

GENERAL SYSTEM AND TERMINAL PROBLEMS

1. Phase 1

Problems associated with:

a. Dial up use - Summary

Period of use: 9/4 - 12/23 for a total of 38 days

Problem	Number of occurrences and/or time involved
Terminal die	56 times
Line noise	7 times
Down a. m.	11 hours, 14 times
Down p. m.	9 hours, 9 times
Fuse blew at NLM	2 times
All lines busy	1 time
2 NLM terminals busy	6 times
Terminal down for repair	1 time
Average delay per search when noted to be unusually slow in response	25 minutes, 3 times
Interruption/line use	2 times
Hardware error	3 times
Improper use (repeated)	1 time
NLM room locked	1 time

b. Indexing

Some indexing of documents appeared to be inconsistent. Concepts were missed as tı ere appeared to be a tendency to index documents only under terms found in text without regard to concepts involved. No pattern could be found for choice of general rather than specific term or vice versa.

Also, the indexing policy followed (noninclusion of narrow terms in broader terms) produces many cases in which the broader term carries fewer postings than the narrower terms.

c. Thesaurus

Choosing a broader or narrower term for a particular listing proves to be unusually difficult since a term may be listed both as a broader term and a narrower term. Also, a particular term may be listed six or more times as relating to various other terms. This organization proves extremely confusing. For instance, in attempting to list all broader or narrower terms for a particular listing, one may easily find the same term under both categories.

d. Bibliographic information

In many cases it proved to be almost impossible to obtain certain bibliographic information for particular documents. Most often topic tags posed the problem; however, at times, country codes, subject codes, etc. proved to be difficult to obtain.

e. Form B (list of relevant documents)

There were cases in which the intelligence analyst was unable to list 10 documents he knew to be relevant prior to the search. In all 30 cases at least two documents were cited. Problems arose in obtaining copies of these as many were not CIRC documents and had to be obtained from outside sources. These then had to be indexed in order to provide the necessary descriptors for system simulation. These problems contributed greatly to unforeseen time delays. Even in cases where more than two documents were cited some of those documents were controlled dissemination and therefore had to be eliminated from consideration.

f. Unexplained problems

In one particular case, four CIRCOL documents were identified which all contained a word (omegatron) in their title. The word (omegatron) was found to be listed in the CIRCOL Dictionary and therefore should be a legitimate word. When searching by the term, omegatron, two different responses occur. At times, 0 documents qualify and at other times, one document qualifies. The one document turns out to be a dummy. CIRCOL personnel were unable to explain why the four previously mentioned documents were not retrieved.

In another case when searching under air & air(+1) & missile (+1), or air-to-air missile, documents were retrieved that did not contain air followed by air followed by missile. CIRCOL personnel determined that perhaps what occurs is that when searching by word distance the same term may not be used more than once, or that not more than two terms at a time in combination will be searched and that in this case, a random choice of first and second or second and third terms may be searched.

At present the answer to both questions remains vague.

In doing printout analysis for determining total retrieval figures, many cases were discovered in which no reason for a particular document's retrieval could be found. In these cases, the words contained in the query were not present in either the title, abstract or descriptor sections of the document. In some cases a related word could be found such as company when searched by corporation, etc.

2. Phase 2

Problems associated with:

a. Dial up use - Summary

Period for use: 2/17 - 5/28 for a total of 70 days

Problem	Frequency of occurrence
Terminal die/disconnect	69
Line noise	26
System down more than 20 minutes	17
System down all A.M. or P.M.	13
Terminal down for repair (A.M. or P.M.)	5
Query response time 15-30 minutes	10
Query response time more than 30 minutes	12
Blank response for qualifying documents	42
Number of documents different for same query when run again	28
Error messages[15]	
Communication problem	25
Internal processing error	19
Unable to process	39

b. Irregular responses of CIRCOL - Examples are:

- Keyword not in dictionary: Although the entry of a keyword term which is not in the dictionary results in a message telling the searcher that that term was not found, what the searcher may not know is whether or not other terms were affected. It appears that in an "&" string, all remaining terms in that string are dropped from the search query by CIRCOL.

[15] When terminal input correct and no identifiable line noise problem.

In a "," string, the remaining terms are not affected. However, this distinction may not be consistent.

In order to obtain a proper response from CIRCOL when a keyword-not-found message is received, the searcher really should reenter the query without the nonfound term.

- Number of qualifying documents varies for same search: In a number of cases, when an identical query was reentered, the response of number of qualifying documents was not the same as previously. Although in some cases this discrepancy could be due to any of the other irregularities described here, or to line noise, or to an intervening update; it also happened when none of these was the case.

This problem became sufficiently apparent about halfway through running queries for Part A that it was decided to run every query twice to check for validity of retrieval number response. Also in rerunning 24 of the 30 Phase 1 searches it was discovered that six resulted in substantially different numbers.

- Narrower search query gives more qualifying documents than a broader one: This discrepancy became most noticeable during the running of Part E queries, where initial queries were expanded to include related and equivalent terms in "or" relationships. Each successive expansion of the original query should have resulted in the retrieval of the same or larger number of documents. This was not always the result, and in the case of one search it was impossible to obtain a reasonable number. That search was therefore dropped from the sample.

- Order of logic in query: The order in which the lines of logic were entered in the query affected the resulting number of qualifying documents. In the case of a long search, the logic ordering could make the difference between having the query run or not run because of exceeding the maximum number of keyword statements. In either case the number of statements and final logic were the same, and only the order of lines was changed.

- Varying maximum query length: Queries could encounter "maximum number of keyword statements exceeded" statements when number of lines was as low as 12. For other searches, queries as long as 23 lines would process with no error message.

- Relational strings: Although we had been warned that relational strings of more than two terms did not process as single strings, methods of building up the longer relational strings gave varying results in terms of number of documents retrieved. Another discrepancy was noted in that using + (plus) relationships or using a reversed string with - (minus) relationships sometimes resulted in a different number of documents being retrieved for the same multiword term.

- Searching on line label: Occasionally CIRCOL would appear to search on line labels as if they were terms. There were only two situations in which it was possible to tell when this had occurred. The first was when an entire string (line) of logic was eliminated because the first term in that string was not found in the dictionary (see above); if the line label (for example "L9") was also not a dictionary term, then a message would sometimes be received stating "L9 not found in dictionary". However, if the label was a term in the dictionary (such as L1, L2, L5, L7, L8, L12, L13) then there was a strong possibility that CIRCOL was searching the label as if it were a term. The second case was identifiable when printouts were ordered. One query had retrieved documents on both the search terms of the string and on the label for that line.

- Won't retrieve on a dictionary term: Although this situation was identified only twice, a question remains as to whether it may also have occurred at other times. One term, "omegation", always failed to retrieve any documents. The second term "cooperation", when qualified by USSR, failed to retrieve any documents for a period of more than a week. This combination had been retrieving documents earlier in the study. CIRCOL personnel were able to correct nonretrieval after being notified.

- Excessive turn-around time: The cause of this problem was never identified. It did not necessarily relate to length of query, number of users on-line, nor periods of time when users were informed to expect slow response. Occasional queries would take longer than 90 minutes, and as there is no method of abandoning a given query short of disconnecting the terminal, nor of knowing whether CIRCOL is actually still processing the query, nor of getting a message to or from the proctor during this time, this problem resulted in considerable irritation and time loss.

APPENDIX IV

FORMS

1. Phase 1

INTELLIGENCE ANALYST
INFORMATION FORM

WESTAT

____________________ DATE ____________________

PURPOSE OF THE EXPERIMENT

A new search technique is being investigated by FTD to provide you an improved search capability. Westat Research, Inc., has been asked to perform an evaluation of this new system. It is felt imperative that this evaluation incorporate real search questions and that evaluation results are based on your assessment of whether documents answer your search question. Furthermore, we need to know which documents are found useful by you (whether or not they actually answer your search question).

I. PRIOR TO SEARCH

A. Please record below a comprehensive statement of your information need IN YOUR OWN WORDS

____

____

____

____

____

____

____

____

____

____

Last Name ____________________ Date ____________________

B. Please list below any relevant documents that you feel answer the above statement of your information need. We need a minimum of at least 5 such documents, including, if possible, ones brought to your attention outside of the CIRC retrieval or profile systems.

1. Author(s) ____________________

Title ____________________

____

Number ____________________

Other information ____________________

2. Author(s) ____________________

Title ____________________

____

Number ____________________

Other information ____________________

3. Author(s) ____________________

Title ____________________

____

Number ____________________

Other information ____________________

4. Author(s) ____________________

Title ____________________

____

Number ____________________

Other information ____________________

5. Author(s) ____________________

Title ____________________

____

Number ____________________

Other information ____________________

Last Name ______________________ Date ______________________

B. Cont.

6. Author(s) ______________________

Title ______________________

____

Number ______________________

Other information ______________________

7. Author(s) ______________________

Title ______________________

____

Number ______________________

Other information ______________________

8. Author(s) ______________________

Title ______________________

____

Number ______________________

Other information ______________________

9. Author(s) ______________________

Title ______________________

____

Number ______________________

Other information ______________________

10. Author(s) ______________________

Title ______________________

____

Number ______________________

Other information ______________________

C. Give the comprehensive statement of information need (#1 above) to the search analyst.

D. Keep the list of known answers to the statement of search needs (#2 above) to submit later.

II. FOLLOWING THE SEARCH

1. The search printout is sent to you in duplicate and you should receive the full text of the retrieved documents.

   Examine the full text of the retrieved documents and the previously known relevant documents. Please record on the extra copy of the computer printout whether or not each document answers the comprehensive statement of your information need (#1 above) by recording a capital "R" next to the right of the listed document.

   Next, please record a capital "U" to the right of the listed document to indicate documents that are found useful to you. [Note that some documents may be relevant but not useful and vice versa.]

2. Please submit to the monitor the list of prior known relevant documents that answered your statement of information need (#IB above) and the marked computer printout stating relevance and usefulness of retrieved documents.

THANK YOU

Westat Research, Inc.
May, 1970

Data Gathering Guide

FTD Search Analyst: Please follow the sequence outlined below in gathering the necessary search information.

A. Prior to Search

1. Identify an appropriate request in light of the purpose of the study.
2. Arrange to have a duplicate copy of the search printout made.
3. Obtain a comprehensive statement of the information need from the analyst, written using his own choice of words.

B. During Search

1. Formulate and conduct the search in the usual manner.
2. Obtain two copies of the dialog printout of the entire search (on-line and off-line portions).
3. Screen for all citations listed on the computer printout and judge them for relevance (R for relevant or NR for not relevant) with regard to the intelligence analyst's comprehensive statement (#1). Mark these on both copies of the printout to the left of each document.
4. Keep the comprehensive statement of search needs and copy of dialog printout until the steps below are completed.

C. Following Search

1. Obtain a copy of the full text of the retrieved documents from the project monitor.
2. Indicate on each document whether the document answers the analyst's comprehensive information statement by an "R" for relevant and "NR" for not relevant to information statement.
3. Indicate on each document whether the document truly answers your search query by a "P" for those pertinent to your query and "NP" for those not pertinent to your query. [Note that some documents may be relevant to the information statement but not pertinent to your query even though retrieved and vice versa.]
4. Please forward the marked documents to the project monitor.

Westat Research, Inc.
May, 1970

Data Gathering Guide

FTD Project Monitor: Please follow the sequence outlines below in gathering the necessary search information.

A. General Procedure

The principal purpose of this portion of the evaluation is to obtain an assessment of relevance from the intelligence analyst and from the search analyst so that we can use these documents as a basis for search simulation.

The intelligence analyst will be asked to formulate and record a comprehensive statement of his information need in his own words. He will then be asked to identify five or more documents that answer this information need prior to the search. He will keep this list until the search is completed.

A search analyst will then conduct the search in the normal manner except that both the dialog printout and off-line printout will be in duplicate for our analysis. The search analyst will be asked to judge each of the retrieved documents for relevance to the statement of information need. ALL of the final retrieved documents will be reproduced and sent to the intelligence analyst with regard to usefulness. He will record his judgments on the duplicate printout and return them along with the list of relevant documents identified prior to the search.

The search analyst will then be given a set of the reproduced retrieved documents and a reproduced set of documents judged to be relevant prior to search by the intelligence analyst. He will judge these documents with regard to relevance to the statement of need and with regard to pertinence to his search query.

These results, along with those above, will then be sent to Westat for further search simulation and evaluation.

B. Prior to Search

1. Arrange for the FTD search analyst to set up the search. (See Data Gathering Guide - FTD Search Analyst.)
2. Provide an information form to the intelligence analyst.

3. Have the intelligence analyst formulate his information needs in his own words.

4. The intelligence analyst will indicate five or more documents that he knows are relevant to his statement of information needs.

5. The intelligence analyst will forward his statement of information needs to the search analyst.

C. During Search

1. The search analyst will formulate and conduct searches in the normal manner except:

   a. Have the dialog printout of the entire search printed in duplicate.

   b. Obtain the off-line printout in duplicate.

   c. Reproduce TWO copies of the full text of the retrieved documents.

2. Have the search analyst indicate relevance of retrieved documents in the left hand margin of the off-line printout.

3. Send printout and duplicate to the intelligence analyst.

4. Provide copies of the full text of retrieved documents to the intelligence analyst.

D. Following Search

1. Have the intelligence analyst indicate (a) relevance of the retrieved documents to their statement of search needs and (b) usefulness of these documents.

2. Provide to the search analyst the full text of retrieved documents and the documents indicated by the intelligence analysts to be relevant.

3. Have the search analysts indicate (a) relevance to the statement of information needs and (b) pertinence to their search query.

2. Phase 2

INTRODUCTION - ANALYST EXPERIENCE AND USE OF RELEVANT DOCUMENTS

This experiment is designed to isolate the effect of two variables on search performance. The first of these variables is the technical experience of the analyst and the second is the use of relevant documents during the query formulation process. Performance will be measured in terms of a recall ratio (the proportion of relevant documents retrieved) and the total number of documents retrieved. The level of favorable performance is assumed to increase as recall is maximized and the number of documents retrieved is minimized. Thus, the "best" search query would be that which exhibits the highest possible recall ratio with the lowest possible number of documents retrieved. This goal should be kept in mind throughout the duration of the experiment.

A group meeting will be held at which time five fairly narrow topics that are agreeable to all concerned will be chosen. Also at this time the group will be asked to identify six documents or possibly more (presently in the CIRCOL system) that are known to be relevant to each topic.

Following the group meeting it is hoped that each analyst will work independently of other group members. Basically, each analyst will be asked to formulate one query for each of the five topics in the usual manner. Upon completion of these five queries, each analyst will be asked to use the pair of relevant documents that will have been provided to him as a search aid in order to formulate a second query for each of the five topics. Thus, each analyst will formulate a total of ten queries, two for each topic.

In both cases text searching will be used as opposed to the usual descriptor (topic tag) searching. Thus, the natural language terms chosen for each query should be based on the terminology used in actual abstracts rather than on topic tag terms.

Group Meeting:

1. Choose five topics that are fair to all analysts.
2. Each analyst specify relevant documents for each topic at the meeting if possible. If not, find relevant documents later and notify Westat as soon as possible. These prespecified relevant documents should be available through CIRCOL.

General Instructions: (For each topic)

1. Formulate a search query using natural language terms combined in the standard Boolean logic and record on the form provided.

2. Open packet containing the appropriate pair of relevant documents and again formulate a search query using natural language terms in Boolean logic; however this time using this pair of documents as a search aid (i.e., possibly adding concepts or terms previously overlooked). Record this query on the form provided.

QUERY RECORD

Date ______________

TOPIC 1

Analyst's Name ________________________

1. Natural language query ______________________________________

____

____

____

____

____

____

____

____

2. Natural language query using 2 prespecified
relevant documents ____

____

____

____

____

____

____

____

____

QUERY RECORD

Date ____________

TOPIC 2

Analyst's Name ____________________

1. Natural language query ________________________________

____

____

____

____

____

____

____

____

2. Natural language query using 2 prespecified relevant documents ______________________

____

____

____

____

____

____

____

____

QUERY RECORD

Date ______________

TOPIC 3

Analyst's Name _______________________

1. Natural language query ____________________________________

____

____

____

____

____

____

____

____

2. Natural language query using 2 prespecified relevant documents ______________________________________

____

____

____

____

____

____

____

____

QUERY RECORD

Date ______________

TOPIC 4

Analyst's Name ________________________

1. Natural language query ________________________________

____

____

____

____

____

____

____

____

2. Natural language query using 2 prespecified relevant documents ______________________________

____

____

____

____

____

____

____

____

QUERY RECORD

Date ______________

TOPIC 5

Analyst's Name ________________________

1. Natural language query ________________________

2. Natural language query using 2 prespecified relevant documents ________________________